VŠB – Technická univerzita Ostrava

Fakulta elektrotechniky a informatiky

ELPOD: Komplexní SEO analyzátor

ELPOD: Complex SEO analyzer

2012 Jiří Baldik

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání bakalářské práce

Student: **Jiří Baldik**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 2612R025 Informatika a výpočetní technika

Téma: ELPOD: Komplexní SEO analyzátor
ELPOD: Complex SEO Analyzer

Zásady pro vypracování:

1. Nastudujte problematiku SEO optimalizace webových stránek.
2. Analyzujte možnosti a přístupy optimalizace on-page faktorů internetových stránek.
3. Cílem práce je implementovat nástroj, který bude schopen navštívit danou webovou stránku a provést kompletní průzkum jejího obsahu.
4. Nad získaným obsahem proveďte základní analýzu kvality zdrojového kódu, copywriting a struktury projektu.
5. Na základě získaných dat proveďte doporučení na optimalizaci, změnu nebo doplnění zdrojového kódu stránky tak, aby byla efektivně indexována vyhledávači.
6. Zkombinujte vaše řešení s již existujícím nástrojem pro SEO optimalizaci a analýzu konkurenčního trhu. Pokuste se porovnat výsledky ostatních nástrojů a případně navrhnout zlepšení na vybraných stránkách.
7. Výsledné postupy a algoritmy dobře okomentujte a pokuste se určit jejich spolehlivost.

Seznam doporučené odborné literatury:

[1] KUBÍČEK, Michal. Velký průvodce SEO : Jak dosáhnout nejlepších pozic ve vyhledávačích. Vydání první. Brno : Computer Press a. s., 2008. 318 s. ISBN 978-80-251-2195-5.
[2] KUBÍČEK, Michal; LINHART, Jan. 333 tipů a triků pro SEO. Vydání první. Brno : Computer Press a. s., 2010. 262 s. ISBN 978-80-251-2468-0.
[3] GRAPPONE, Jennifer; COUZIN, Gradiva. SEO - Search Engine Optimalization. Překlad: Roman Skřivánek, Dana Balaštíková. Vydání první. Brno : ZONER software, s.r.o., 2007. 328 s. ISBN 978-80-86815-85-5.
[4] JANOUCH, Viktor. Internetový marketing : Prosaďte se na webu a sociálních sítích. Vydání první. Brno : Computer Press a. s., 2010. 304 s. EAN: 9788025127957.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Radoslav Fasuga, Ph.D.**

Datum zadání: 18.11.2011

Datum odevzdání: 04.05.2012

doc. Dr. Ing. Eduard Sojka
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlášení o autorství

Prohlašují, že tuto bakalářskou práci jsem vypracoval samostatně. V práci jsou uvedeny veškeré literární prameny a publikace, ze kterých jsem čerpal.

Ve Studénce dne: 1. 5. 2012 Podpis:

Poděkování

Na tomto místě bych chtěl poděkovat svému vedoucímu Ing. Radoslavovi Fasugovi, Ph.D., za velmi dobré vedení bakalářské práce, bez níž by nevznikla.

Abstrakt

Práce obsahuje základní seznámení s problematikou SEO a popis implementace aplikace pro porovnávání domén z hlediska SEO kvality zdrojového kódu. První část práce obsahuje základní informace o tom, jak pracují internetové vyhledávače a jejich roboti. Tedy jak vyhledávače procházejí stránky, co je na těchto stránkách zajímá a jak určují pořadí stránek ve výsledku vyhledávání. Z těchto faktorů je také nastíněna volba klíčových slov a jejich umístění na stránce, včetně objasnění důležitosti zpětných odkazů. V další části práce, jsou obsaženy informace, co dělat pro to, aby byly internetové stránky pro SEO robota atraktivní a aby se na stránky rád vracel. To zahrnuje napsání zdrojového kódu stránky, umístění textu na stránce či umístění klíčových slov na stránce. Poslední část obsahuje popis implementace aplikace a použitých technologií. Do implementace se řadí zejména diagramy znázorňující chod aplikace, ať už z hlediska programového, či databázového.

Klíčová slova

SEO analyzátor, on-page faktory, PHP, MySQL, stahování stránek, parsování stránek, vytváření n-gramů, porovnávání n-gramů, podobnost domén, podobnost stránek, podobnost klíčových slov

Abstract

Content of this work are a basic rules of SEO issues and the description of implementation of application for comparison of domains in terms of SEO quality of source code. First works section include basic information about how works the internet searchers and it´s robots. So as searchers read pages, what interest them on these pages and how they determine the order of a search results. From these factors is also outlined importance of choosing key words and their location on the webpage, include clarifying the importance of back links. In the next part of the work are clarify information, what to do for websites to do them attractive for SEO robots and to make SEO robot return often. This includes writing of source code of webpage, location of text on webpage or location of key words on webpage. Last part of this work includes description of applications implementation and used technologies. Implementation particularly includes diagrams, which shows running of application in program view or database view.

Key words

SEO analyzer, on-page factors, PHP, MySQL, downloading of webpages, parsing of webpages, creating of n-grams, comparing of n-grams, similarity of domains, similarity of key words

Seznam použitých symbolů a zkratek

SŘBD – systém řízení báze dat
MySQL – typ SŘBD
IS – informační systém
SEO – Search Engine Optimalization (optimalizace pro vyhledávače)
PHP – scriptovací programovací jazyk, pracující na straně serveru, slouží k vytváření dynamických stránek
PPC – Pay Per Click (platba za klik)
HTML - HyperText Markup Language (hypertextový značkovací jazyk), jazyk pro vytváření základní obsahové kostry webové stránky
on-page – faktory webové stránky ovlivňující umístění stránky ve výsledcích vyhledávání, tyto faktory lze ovlivnit editací zdrojového kódu stránky
off-page – faktory webové stránky ovlivňující umístění stránky ve výsledcích vyhledávání, obsahují například odkazy směřující na danou stránku z jiných stránek

Seznam obsažených obrázků a tabulek

Obsah

1. Úvod

Cílem této práce je vytvořit komplexní SEO analyzátor, který bude schopný analyzovat webové stránky v českém jazyce, napsané v jazyce HTML. Program bude sloužit převážně pro porovnávání konkurenčních webů, ale může se využít například pro vyhledávání podobných stránek na vlastním webu. Aplikace by se měla přibližovat již existujícím aplikacím, jako je například spyfu.com.

Při zadávání webů pro porovnání, si uživatel navolí, zda chce procházet domény pouze 2. řádu nebo i nižšího řádu. Dále si může zvolit, které pravidla pro roboty, bude aplikace respektovat, tedy jako který robot se bude aplikace chovat (např. Googlebot). Pravidla pro všechny roboty (*), jsou respektována vždy.

Pro získání výsledků program prochází zadané weby, stáhne všechny podstránky daného webu a uloží je do databáze. Z těchto podstránek vytřídí pouze text, s kterým bude dále pracovat, což jsou obsahy atributů nebo elementů ovlivňující kvalitu stránky, z pohledu SEO. Tento text je také uložen do databáze. Následně z tohoto textu program vytvoří klíčové slova, tedy n-gramy a uloží je do tzv. slovníků v databázi. Tyto slovníky obsahují různé kombinace klíčových slov, a unikátní klíč k danému slovu. Program poté nemusí pracovat s textem, ale vystačí si pouze s odkazy na dané klíčové slova, což rapidně zvýší rychlost programu.

U uložených n-gramů se kontroluje, ve kterém elementu nebo atributu se na stránce nachází (důležitost jednotlivých elementů je odlišná). Podle toho kde se nachází, je výskyt n-gramu na stránce pře násoben hodnotou, kterou si nastaví uživatel. Tímto se zvýší počet výskytu daného n-gramu, což ovlivňuje důležitost.

Takto vytvořené a převážené n-gramy, pro každou stránku zvlášť, budou porovnávány s n-gramy z jiných stránek. Před samotným porovnáním program profiltruje stránky vůči sobě, aby zjistil, které stránky si jsou podobné a které ne. Tohle profiltrování, ušetří spoustu času, jelikož program nemusí zbytečně počítat podobnost klíčových slov u dvou stránek, které si nejsou podobné. Propustnost filtru, tedy minimální schodu v procentech nebo v množství stejných slov, si uživatel nastaví sám. Výsledek filtru je před výpočtem nabídnut uživateli pro kontrolu, kde má možnost manuálně nastavit, které stránky jsou si podobné.

U stránek které prošly filtrem, se přechází k analýze. Uživatel má opět na výběr, který způsob výpočtu podobnosti použít, včetně nastavení penalizací. Výsledkem této analýzy jsou informace, říkající uživateli, které klíčové slova a v jakém množství se vyskytují na jeho a konkurenční stránce. Z těchto informací program uživateli doporučí, co udělat proto, aby se ve vyhledávači dostal před konkurenci, tedy zda přidat nebo ubrat klíčové slova a kde.

2. Problematika SEO

2.1 Co je to SEO?

Zkratka SEO pochází z anglického slovního spojení Search Engine Optimization, tedy v překladu optimalizace pro vyhledávače. Pokaždé když do jakéhokoliv vyhledávače, ať už Google, Seznam nebo jiného, zadáme hledanou frázi, můžeme říct, že jsme se dostali do kontaktu se SEO. Vyhledávač nám jako výsledek hledané fráze nabídne různé stránky v určitém pořadí a právě tyto stránky, jsou výsledkem SEO optimalizace.

Zjednodušeně můžeme říci, že SEO optimalizace se snaží dostat právě naši konkrétní stránku na první místo ve vyhledávači. Je to jakýsi sled pravidel, podle kterých se mají weby, které se chtějí optimalizovat na určitou frázi řídit. Ne vždy to však stačí, pokud máme například perfektně zoptimalizovanou stránku, ale konkurence si u daného vyhledávače platí reklamu, neboli pevně dané umístění na konkrétní dotaz, nenaděláme nic, leda si také zaplatit reklamu. Nezáleží tedy jen na tom, jak moc kvalitně máme stránku napsanou, této konkrétní části SEO se říká on-page faktor, ale také záleží na tzv. off-page faktorech, které nám zajišťují právě zmiňované reklamy, nebo odkazy na danou stránku.

2.2 Jak fungují vyhledávače

2.2.1 Rozdíl mezi katalogem a fulltextovým vyhledávačem

Co je to katalog?

Katalog je v podstatě web, obsahující různé kategorie a podkategorie, ve kterých jsou umístěny odkazy na hledané stránky. Takovýmto katalogem je například u nás asi nejznámější stránka Firmy.cz od firmy Seznam.cz, nebo světově nejznámější katalog Yahoo!. Jedním z nejkvalitnějších katalogů je ovšem projekt ODP (Open Directory Project) – dmoz.org, který spravují dobrovolní editoři z celého světa.

Oproti vyhledávačům se do katalogů stránky musí registrovat, tedy vkládat ručně. Uživatelé většinou zadají adresu webové stránky, k ní krátký popis a pár klíčových slov charakterizující tuto stránku. Takto zaregistrovaná stránka obvykle projde schválením a po té je zpřístupněna nejen uživatelům webu, ale také právě vyhledávačům. Ano! Vyhledávače často považují katalogy za spolehlivé zdroje informací a není tedy divu, že při mnoha dotazích se právě katalogy vyskytují na prvním místě ve vyhledávači. Například právě katalog ODP, díky své nestrannosti a systému přidávání odkazů do katalogu, jsou většinou stránky tohoto katalogu vyhledávači dobře hodnoceny. Google dokonce považuje ODP za tak kvalitní katalog, že pokud na daném webu nenajde dostatek informací pro, popsaní stránky v odkazu, použije popis z katalogu ODP.

Katalogy samozřejmě dnes obsahují i vyhledávací funkce. Katalog vyhledává stránky ve své databázi, podle výše zmiňovaného popisu, klíčových slov, titulu stránky nebo jiných zdrojích informací, které jste při registraci svého webu do katalogu vyplnil, nebo byly vyplněny editorem.

Co může být značnou nevýhodou katalogů, je fakt že většina používaných katalogů nabízí koupení si pozice zobrazovaní. Katalog často nezajímá obsah stránky, kterou má uloženou v databázi, obvykle jej ani nezná. Editoři sice občas můžou stránky procházet, ale není to pravidlem, proto si může každý nově registrovaný web zadat jakýkoliv popis, klíčová slova nebo titul. To má za následek, že při vyhledávání nového automobilu, se nakonec dostaneme na dovolenou do Karibiku. Tímto ovšem nechci naznačit, že by byla registrace do takovýchto katalogů zbytečná, naopak je dobrá pro získání tzv. zpětných odkazů, ke kterým se vrátím v pozdější kapitole, které můžou pozitivně ovlivnit off-page faktor stránky.

Jaký je tedy rozdíl mezi katalogem a fulltextovým vyhledávačem?

Jako hlavní rozdíl bych označil, že vyhledávače se nesnaží pojmout co nejvíce odkazů na webové stránky bez ohledu na to, zda tyto odkazy splňují svůj popis, ale snaží se především o to, aby na zadaný dotaz byly vybrány přesně ty stránky, které tomu svou kvalitou i obsahem odpovídají. Ovšem nemůžu říci, že by všechny katalogy ignorovaly obsah stránek, například v případě zmiňovaného ODP to určitě neplatí.

Dalším zásadním rozdílem je vyhledávání v databázi. Oproti katalogům se do vyhledávačů stránky neregistrují (do některých ano), ale stránky jsou v síti nalezeny a zpracovány. Zpracováním takto nalezených stránek, si vyhledávač zjistí, o čem je obsah nalezené stránky, vygeneruje si vlastní klíčové slova k této stránce a podle takto získaných informací může uživateli nabídnout rozhodně relevantnější výsledek, nežli katalog.

Za vyzdvižení dále stojí, že v současné době, stejně jako katalogy i vyhledávače nabízejí možnost placeného umístění, tedy jakési reklamy. Ovšem musím zdůraznit, že doposud platí, že takto zaplacené odkazy jsou zvýrazněny a tedy uživatele dostatečně informují o tom, že tohle zrovna nemusí být stránka, kterou zrovna potřebuje. Otázkou zůstává, kolik uživatelů tohle bere v potaz?

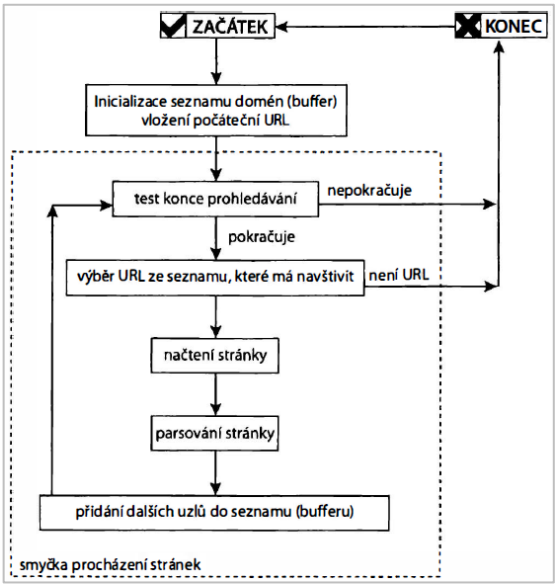
Tohle jsou pouze základní rozdíly mezi katalogem a vyhledávačem. V dalších kapitolách se budu podrobněji zabývat tím, jak fungují vyhledávače, což vystaví na světlo další z mnoha rozdílů mezi katalogem a vyhledávačem.

2.2.2 Pavouci na webu

Pro to aby vyhledávače zajišťovaly aktuální výsledky a měly v databázi, co nejvíce aktuálních stránek slouží tzv. pavouci, nebo taky roboti, červi a jiné podobné názvy. Vyhledávače tyto pavouky nechávají neustále procházet celý Internet a aktualizovat výsledky ve své vlastní databázi. To umožňuje rychlé zobrazení výsledků, takže při zadání dotazu nečekáme na výsledek několik dní, ale jen pár milisekund. Pavouci Googlu umí v největší zátěži stahovat až sto stránek za jedinou sekundu, takže průběžně tak procházejí miliardy stránek.

Jak pracují pavouci

Pavouk svou cestu začne tím, že si z databáze načte odkazy stránek, které bude procházet. Začíná od hlavní stránky, kde si pročte její obsah a uloží sis jej do skladiště. Z tohoto obsahu přečte další odkazy, obsažené v elementu href párových značek <a>. Takto získané odkazy prochází tak dlouho, dokud kompletně neprojde celý web (Obrázek 1). Takovýchto pavouků je hned několik a každý plní svou funkci. Jeden stahuje odkazy, druhý vyhledává obrázky, třetí prochází PDF dokumenty, další stahují hlavičky dokumentu, kontrolují změnu stránky, čas stažení stránky atd. [3]



Obrázek 1: Jak pracují pavouci

2.2.3 Jak přistupuje vyhledávač k novým stránkám

Při první návštěvě stránky vyhledávačem je relativně jednoduché nechat si zabanovat stránku. A to tím způsobem, že na daném odkazu, který vyhledávač navštíví, se nic nenachází. Jakmile tohle vyhledávač zjistí, párkrát se ještě pokusí o navázání spojení a pokud stránka stále nereaguje, tak si tento odkaz uloží do databáze nedostupných odkazů, tedy odkazů, které nebude navštěvovat. V praxi to ovšem neznamená, že jej nenavštíví vůbec, ale například jen jednou měsíčně, aby se ujistil, zda je odkaz stále nedostupný. Pro takovýto nedostupný odkaz, je docela těžké se opět dostat do vysokých míst ve vyhledávači.

Při úspěšném stažení stránky, vyhledávač přechází k dalším krokům, které budou vysvětleny níže. Pro nastínění, je to například výše zmiňované čtení odkazů, převedení velkých písmen v odkazech na malé, nahrazení znaků jako je například tilda, za ekvivalentní znak, který se může vyskytovat v URL adrese, převedení relativních odkazů na absolutní, dalšími úkony je například parsování důležitých tagů a čtení jejich obsahů, atd.

2.2.4 Jak pracuje Google

Uvedu stručnou, ale výstižnou a jednoduchou citaci z knihy Velký průvodce SEO od autora Michala Kubíčka:

> *„Google se skládá ze serveru obsahujícího seznam URL adres, které pravidelně zasílá těmto pavoukům. Stažené stránky jsou posílány do skladového serveru (storeserver). Ve skladišti se stránky komprimují a uloží dále do depozitáře. Každá stránka dostane unikátní identifikační číslo, kterému se u Google říká docID. O zařazení do rejstříku (indexu) neboli indexování se stará tzv. indexer a sorter (třídič). Dalo by se říct, že indexer je srdcem vyhledávače. Obstarává totiž celou řadu důležitých činností.“*

Jak pracuje Google indexer?

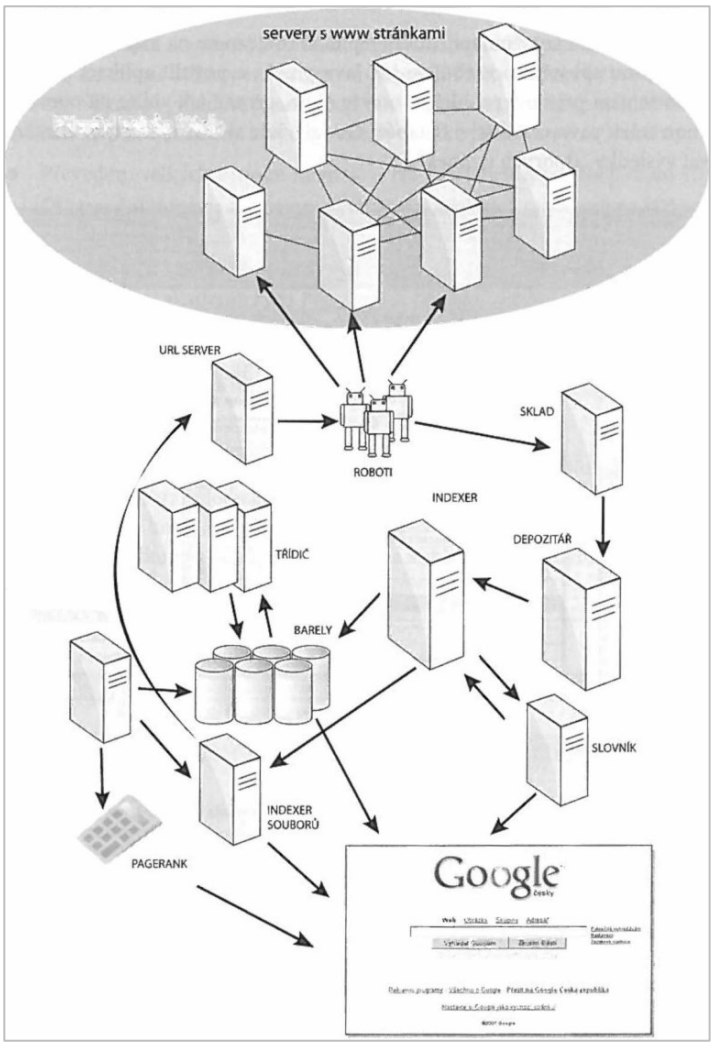
Pro úvod, seznam úkonů, které indexer mimo jiné provádí (Obrázek 2):

- Čte informace z depozitáře
- Dekomprimuje komprimované stránky
- Analyzuje rozložené stránky
- Komunikuje se zásobárnou stránek
- Spolupracuje se slovníkem
- Extrahuje odkazy a adresy URL

Googlebot předává indexeru celý text stažené stránky. Tyto stránky jsou uloženy v indexové databázi Googlu. Google index je seřazen abecedně, podle hledaného výrazu. Každá vyhledávací fráze, tedy klíčové slovo, obsahuje odkaz na dokument a pozici, na které se v daném dokumentu nachází.

Pro zvýšení výkonu indexer čistí obsah dokumentu. Z dokumentu odstraní stop slova (např. české stop slova: a, je, v, na, ale, apod.) a některé jednotkové číslice a samostatné písmena. Google také ignoruje interpunkce a mnohonásobný výskyt mezer vedle sebe. Nakonec převede všechny znaky na malá písmena, pro lepší zpracování.

Dokument je převeden na množství tzv. barelů. Každé slovo, je převedeno na wordID a zapsáno do slovníku. Pokud se v dokumentu vyskytne stejné slovo víc než jednou, není mu znova přiřazeno nové wordID, ale je uloženo do tzv. hit listu. Hit list obsahuje wordID a informace o slovu jako je jeho pozice, velikost fontu, řez, apod. Tyto hity jsou odeslány do barelů. Nakonec je lokální slovník obsahující seznam slov vyskytujících se v dokumentu sdílen s globálním slovníkem Googlu, do kterého jsou zapsány všechny nové slova, které se v daném dokumentu objevily. Globální slovník Googlu obsahuje kolem čtrnácti miliónu slov.



Obrázek 2: Jak pracuje Google

Indexer zajišťuje další důležitou funkci, kterou je extrahování URL adres a anchor textu ze stránky. URL adresy extrahuje z atributu href u párového tagu <a>. Anchor text je tedy obsah vyskytující se mezi těmito párovými znaky, tedy text, který uživatel vidí jako odkazující. Následně na řadu přichází další program zvaný URLresolver, který převádí relativní adresy na absolutní. Anchor textu je přiřazeno docID, tedy id dokumentu, na který daný anchor text (tedy i odkaz) odkazuje. Poté tyto odkazy páruje se stránkami, které ve zpracovaném rejstříku. URLresolver navíc vytváří databázi odkazů, ze které poté čerpá další program pro výpočet PageRanku. Adresy jsou seřazeny podle docID. [7][9]

Za zmínku stojí, že Google neposkytuje těmto datům jen jediný server, ale data jsou rozdělena do řady oddělených serveru, tzv. *Google data center*. Dotazy pro vyhledávání jsou pak distribuovány přes hlavní Google server do těchto data center.

2.2.5 Známkování stránek vyhledávači

Každý vyhledávač má vlastní hodnotící systém, podle kterého určuje pozici zobrazení dané stránky ve výsledku. Tento hodnotící systém je ovlivňován mnoha hodnoticími faktory, jako například odkazové popularita, významnosti, četnosti hledaného výrazu, aj.

Přesné algoritmy pro výpočet hodnot jednotlivých ranků nejsou známy, jelikož to jsou jedny z nejstřeženějších tajemství vyhledávačů.

PageRank (Google)

PageRank zavedli vývojáři od společnosti Google, aby měli možnost řazení stránek při zobrazení výsledku. PageRank ovšem není ovlivněn, jak by si mnozí mohli myslet, počtem výskytu vyhledávané fráze na určité stránce. Takto to sice kdysi fungovalo, ostatně funguje i dnes ale ne u PageRanku, ale vývojáři museli přijít s jiným algoritmem, protože při vytvoření stránky, obsahující pouze co nejvíce dané fráze nastává problém s relevantností výsledku. Nicméně je třeba dodat, že skutečnost, v jakém množství se vyskytuje daná fráze na stránce, je stále velice důležitým aspektem při hodnocení webu, ne však jediným.

Je třeba zmínit, že podobně jako je tomu u frekvence hledané fráze na webu, tak ani PageRank, není nejdůležitějším a už vůbec ne jediným, faktorem ovlivňující hodnocení stránky vyhledávačem. Například vyhledávač Google má takovýchto faktorů asi dvě stě!

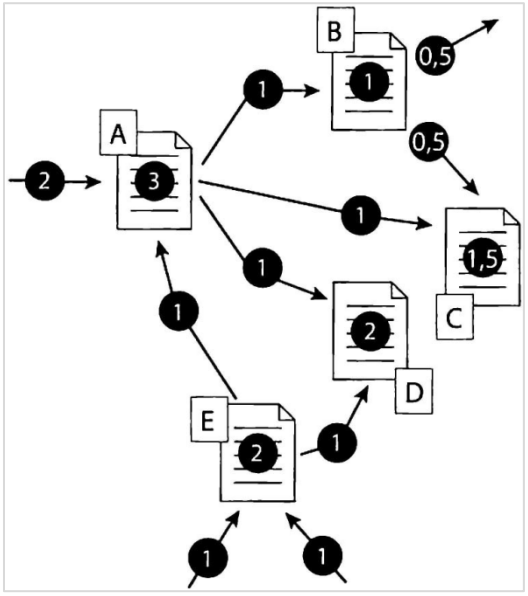
PageRank je tedy hodnotící systém vyhledávače Google. Je to algoritmus, který na jedenáctistupňové škále od 0 do 10 hodnotí hodnověrnost stránky, s tím, že čím vyšší je hodnota PageRanku, tím je hodnocení stránky lepší. PageRank nijak nezávisí na hledaném slově a vlastní PageRank má každá jednotlivá stránka, tedy URL adresa, nikoliv celý web dohromady, neboli doména. Výsledkem PageRanku je relativní hodnota, která se časem mění, může stoupat, ale i klesat. Rovnice na výpočet této relativní hodnoty obsahuje více než 500 miliónu proměnných a dvě miliardy členů!

Jak vlastně PageRank funguje?

Citace myšlenky, ze které vychází algoritmus PageRank, z knihy Velký průvodce SEO, od autora Michala Kubíčka:

> *„Algoritmus PageRank vychází z Kandall-Weiovy teorie hodnocení z padesátých let minulého století, která razí ideu porovnání významu lidí a věcí na základě vlivu, který na sebe vzájemně mají.”.*

PageRank tedy představuje hodnotu důvěryhodnosti, tedy kolik stránek hodnocených podle stejného vzorce na danou stránku odkazuje. Je to algoritmus, který se počítá ze zpětných odkazů na danou stránku, tedy z odkazů na jiných webových stránkách. s tím, že platí pravidlo „čím vyšší má odkazující stránka PageRank, tím vyšší má tento odkaz na mou stránku hodnotu“ (Obrázek 3). Autoři této myšlenky vychází z předpokladu, že pokud je nějaká stránka často citována nebo odkazována v jiných publikacích, tak bude její obsah pravděpodobně kvalitní.



Obrázek 3: Příklad PageRanku

V praxi to funguje tak, že každá stránka předává část své hodnoty PageRanku stránkám, na které odkazuje. Tato předávána hodnota je tedy ovlivněna tím, jakou hodnotu PageRanku má odkazující stránka a na kolik stránek odkazuje. Pokud bude odkazovat stránka s PageRankem 6 na 10 stránek, předá větší hodnotu svého pageRanku odkazovaným stránkám, než stránka s PageRankem 6 odkazující na 100 stránek. [6]

Jsou dvě cesty, jak zjisti hodnotu PageRanku své, nebo prohlížené stránky. Pokud chceme zjistit hodnotu prohlížené stránky, postačí nám program zvaný „Google Toolbar“, který se jednoduše nainstaluje do prohlížeče a zobrazuje nám hodnotu PageRanku prohlížené stránky. Pokud chceme vložit na své stránky obrázek, který ukazuje hodnotu PageRanku této stránky, postačí nám do atributu src u elementu img vložit hodnotu „http://pagerank.yuhu.cz/pr.php?url=www.mojedomena.cz/url“. Tento odkaz nám bude automaticky generovat obrázek s aktuální hodnotou PageRanku dané stránky.

Mýty panující okolo PageRanku:

- Čím vyšší je hodnota PageRanku, tím výše bude stránka umístěna ve výsledcích vyhledávání
- Čím více odkazů vede na stránky, tím vyšší mají PageRank
- Hodnota PageRanku je neměnná, případně jen rostoucí
- Hodnota PageRanku je aktuálním hodnocením, jak vás vnímá Google
- Cílem SEO je vysoký PageRank

S-rank (Seznam)

S-rank je hodnotící systém od českého vyhledávače Seznam, funguje podobně jako PageRank a počítá se téměř stejně. Vysoký S-rank znamená, obdobně jako je tomu u PageRanku, mnoho kvalitních tematických zpětných odkazů na danou stránku. Hodnotí pouze česky a slovensky psané stránky. Výsledná hodnota S-ranku se aktualizuje asi jednou za čtrnáct dní a nabývá hodnot od 0 do 10.

Algoritmus pro výpočet hodnoty S-ranku zohledňuje jednak odkazy, odkazující na danou stránku, ale i to, kam ze stránky odkazy vedou. Je známé, že se S-rank vypočítává váženou nelineární kombinací různých veličin, v nihž výrazně převažují off-page faktory. Při výpočtu se Seznam inspiruje principem Hubs & Authorities. To znamená, že rozlišuje mezi dvěma typy stránek, rozcestníky (hubs) a autority (authorities). Rozcestník je stránka, která odkazuje na mnoho autorit a autorita je stránkou, na kterou odkazuje mnoho rozcestníků. V této souvislosti používaný algoritmus se nazývá HITS, který nakolik je stránka autoritou a na kolik je egocentrická. Vzájemně se obě hodnoty podporují. [8]

2.2.6 Pro které vyhledávače optimalizovat

Internetových vyhledávačů existuje na internetu stovky, u nás se mezi nejpoužívanější řadí Seznam a Google. Donedávna Seznam v podílu vyhledávačů na českém trhu jasně převažoval nad Googlem, ale dnes už to není pravda. Google se pomalu ale jistě, dostává na pozici primárního vyhledávače v Čechách (Tabulka 1). Dá se říci, že Google převážně používají uživatelé, kteří s PC pracují déle a umí s ním lépe zacházet, než uživatelé, kteří si jdou na internet přečíst pouze emaily. Tito uživatelé často nevědí ani o existenci Googlu, někdy dokonce nepoužívají ani adresní políčko svého prohlížeče, ale www adresu zadávají přímo do vyhledávacího políčka Seznamu (jsou to asi 2% uživatelů Seznamu). Pro tyto uživatele často znamená, že co se nenachází na Seznamu, prostě na internetu není. [11][12][13]

	2006	2007	2008	2009	2010	2011
Seznam.cz	62,25%	62,42%	62,83%	56,64%	47%	47,50%
Zbozi.cz	-	-	-	3,25%	5%	3,70%
Firmy.cz	-	-	-	-	5,10%	7,30%
Google	23,67%	26,81%	31,39%	31,61%	43%	47%
Centrum	5,05%	4,12%	2,74%	1,76%	1,10%	1,00%
Atlas	2,62%	2,06%	1,13%	0,52%	-	-

Tabulka 1: Podíl vyhledávačů (2006-2011)

Pro který vyhledávač tedy optimalizovat, Google nebo Seznam? Dá se říci, že si můžeme vybrat podle toho, co chceme na našich stránkách provozovat. Pokud je obsah našich stránek, zaměřen například na technologii Internetu, bude lepší optimalizovat pro Google. Ale pokud je obsah stránek například internetový obchod s parfémy, bude pravděpodobně lepší zaměřit se na Seznam. Důležité ovšem je, nezaměřit se pouze na Google nebo Seznam, ale optimalizovat přinejlepším pro oba vyhledávače.

2.2.7 Spolupráce a komunikace s pavouky

Důležitým krokem, pro zveřejnění našich stránek je, aby nás zaregistrovali pavouci jednotlivých vyhledávačů. Jsou dva způsoby, prvním a jednodušším je, zaregistrovat novou adresu přímo ve vyhledávači (Tabulka 2). Druhým způsobem a pravděpodobně tím kvalitnějším je, mít svou URL adresu umístěnou někde na webu, jak již bylo popisováno výše, aby si nás robot sám našel a zaregistroval do své databáze.

Vyhledávač	Adresa pro registraci
Google.com	*http://www.google.com/addurl/*
Seznam.cz	*http://www.seznam.cz/pridej-stranku*
Jyxo.cz	*http://www.jyxo.cz/d/submit*
MSN.com	*http://search.msn.com/docs/submit.aspx*

Tabulka 2: Adresy pro registraci stránek do vyhledávačů

Zadání URL adresy, prostřednictvím registrace přímo ve vyhledávači nijak neovlivňuje umístění adresy ve výsledku hledání, dokonce to ani nezaručuje, že se nová adresa zobrazí ve výsledcích vyhledávání

Co pavouci na stránkách uvítají

- *Dobrý obsah*
  - Používání slov, která může vyhledávač zahrnout do svého slovníku, resp. Dotazu ve vyhledávání
  - Validní kód HTML
  - Stránka bez obrázků by neměla být větší než 100kb
- *Pěkné adresy*
  - Vyhledávač má raději stejné adresy jako člověk (lépe si je zapamatuje), před adresou http://www.domena.cz/index.php?page=cenik upřednostní adresu http://www.domena.cz/cenik
- *Kvalitní navigaci*
  - Funkční odkaz, jak interní tak externí
  - Každá stránka, kterou chcete, aby pavouk za indexoval, musí být přístupná statickým odkazem, tzn., že stránka, které je přístupná pouze JavaScriptem, AJAXem nebo například vyhledávacím formulářem nebude za indexovaná, protože pavouk ji jednoduše nevidí
  - Dobré je používat mapu webu, aby pavouk mohl navštívit i ty stránky, které například nemají odkaz uvedený nikde v textu
  - Při přidání nových stránek, je dobré vygenerovat novou XML mapu (sitemap), ze které pavouk zjistí, že se na webu něco děje
  - Jednoduchá hierarchie webu, tzn., že každá stránky by měla být dostupná z úvodní stránky, ideálně na jeden, maximálně dva prokliky

- *Aby nebloudili*
    - Při přesunu stránky, nebo změně názvu souboru, nastavovat původní URL, aby přesměrovala uživatele na novou URL
    - Upozornění pavouků, zda se jedná o dočasné nebo trvalé změny
    - Určení kam robot smí, nebo nesmí vstoupit pomocí souboru robots.txt, nastavení hodnotu atributu rel v elementu <a> na nofollow nebo nastavení v meta elementu robots hodnotu nofollow

Jak zakázat robotovi přístup na různé stránky

Může se nám stát, že chceme robotům zakázat přístup na některé naše stránky. V podstatě máme tři možnosti jak to provést:

- První je ta, že se odkaz na danou adresu nebude nacházet nikde na webu. Na tento způsob se nedá moc spoléhat, protože si jen těžko uhlídáme neexistenci jediného odkazu na celém internetu.
- Druhou možností je, že použijeme URL s mnoha parametry v query stringu, kde nám obvykle stačí mít více než 4 parametry. Adresa poté může vypadat například takto http://www.domena.cz/index.php?stranka=auto&znacka=bmw&model=3&karoserie=sedan. Tento způsob je v praxi použitelnější a spolehlivější než předchozí způsob.
- Nejjednodušším způsobem je, jednoduše robotům vstup na danou stránku zakázat. Ovšem na druhou stranu, tento způsob může být také nebezpečný, protože tyto zakázané stránky vyhledávají útočníci, které obsah těchto stránek naopak velice zajímá. Proto je dobré, tyto zakázané stránky chránit například heslem.

Důvodů proč zakázat přístup na stránku je hned několik. Cílem odkazu můžou být například například stránky typu:

- Výsledek vyhledávání
- Interní fórum webu
- Stránky pro tisk
- Placený obsah

K zablokování takovýchto odkazů existuje několik způsobů, mezi ty nejpoužívanější patří:

- Zablokování jednotlivých stránek zvlášť v meta elementu robots v hlavičce HTML dokumentu
- Pomocí atributu rel=““ v elementu <a>
- Pomocí souboru robots.txt umístěném v kořenovém adresáři domény

Meta značky v hlavičce HTML souboru

Omezovat roboty, můžeme v hlavičce HTML soubor, mezi tagy <head></head> v meta tagu <meta name=“robots“ content=“…“ />. Do atributu name, uvedeme robota, kterého chceme omezit (při použití robots, budou omezení všichni roboti) a do atributu content zapíšeme omezovací pravidlo.

Hodnoty atributu content:

- index/noindex – robot smí/nesmí přidat tuto stránku do své databáze
- follow/nofollow – robot smí/nesmí následovat odkazy na této stránce
- all – totéž co “index, follow“ (prázdná hodnota má význam all)
- none – totéž co “noindex, nofollow“

Například příkazem <meta robots=“googlebot“ content=“noindex, follow“ /> říkáme Googlebotu, že tuhle stránku nemá přidávat do své databáze, ale má následovat všechny odkazy na ní uvedené.

Tenhle způsob je ovšem funguje jen částečně, protože robot musí projít a rozparsovat celou stránku, aby nakonec zjístil, že danou stránku nemá indexovat. Navíc danou stránku bude pravidelně navštěvovat, aby zjístil, zda se příkaz nezměnil. Meta element v hlavičce tedy moc práce robotovi neušetří.

Zablokování pomocí atributu rel

Dalším možným řešením je příkaz rel=“nofollow“ v elemetnu <a>, který říká robotovi, že tento link nemá navštěvovat. Tento příkaz můžeme uvádět u odkazů, nad kterýma nemáme dostatečnou kontrolu. Takové odkazy jsou například různé diskuzní fóra, diskuze pod články, apod. Zamezíme tak snížení relevance odkazů a hodnocení dané stránky za odkazovaní na pochybné webové stránky.

Ovšem tenhle způsob také není úplně dokonalý, protože robot sice nebude následovat daný odkaz, případně nepředá hodnotu odkazující stránky, ale cílovou stránku smí nadále indexovat. Navíc ne všechny vyhledávače tuto možnost podporují.

Práce se souborem robots.txt

Odborný název tohoto způsobu je Robot Exclusion Protocol (REP) a v podstatě je to mechanismus, jak předávat robotům informace kam na stránkách smí a kam nesmí. Robots txt je psán ve sledu po sobě jdoucích příkazů s tím, že každý začíná na novém řádku.

Prvním prvkem v tohoto souboru je řádek začínající příkazem “User-agent“. Ten předepisuje, pro kterého robota budou následující pravidla platit. Za ním následují příkazy (libovolné množství, ale minimálně jeden) typu “Disallow“ nebo “Allow“. Disallow zapisuje stránky, na které robot nemá přistupovat a Allow stránky, kde robot může přistupovat. Význam a příklady jednotlivých příkazu viz níže.

Vrátím se ještě k umístění souboru robots.txt. Soubor robots.txt musí být umístěn v kořenovém adresáři domény druhého nebo třetího řádu a musí být zapsán malými písmeny, s tím, že tento soubor platí jen pro aktuální doménu, ve které je umístěn. Takže pokud je cesta k souboru *http://domena.cz/robots.txt*, tak ovlivňuje pouze doménu *http://domena.cz* a možné domény *http://subdomena.domena.cz* nebo *http://www.domena.cz* (pokud je jiný obsah) ovlivňovat nebude. Tohle platí i obráceně.

Záleží také na protokolu HTTP, takže soubor umístěný v *http://www.domena.cz/robots.txt* nebude ovlivňovat adresu *https://www.doman.cz*. Nicméně adresy s protokolem https obvykle vyhledávače neindexují.

Příklady a vysvětlení způsobu zápisu souboru robots.txt

Pravidla aplikovaná pro roboty uvedené v User-agent:

- User-agent: *
  - Následující pravidla budou aplikovaná pro všechny roboty
- User-agent: Googlebot
  - Následující pravidla budou aplikovaná pro robota od vyhledávače Google
- User-agent: Seznambot
  - Následující pravidla budou aplikovaná pro robota od vyhledávače Seznam

Pravidla Disallow aplikovaná pro roboty z předchozího kroku:

- Disallow: /
  - Robot nesmí nikam
- Disallow: /php/
  - Robot nemá přístup do adresáře php
- Disallow: /in
  - Robot nesmí nikam, co začíná slovem in (například index.php)
- Disallow: /*.php$
  - Robot nesmí číst soubory, končící koncovkou .php
  - Ze zápisu je zřejmé, že při určování pravidel pro roboty, můžeme využívat regulárních výrazů, bohužel zde platí, že tento zápis nepodporují všechny vyhledávače

Konstrukce Allow je obdobná jako Disallow, s tím, že můžeme vše zakázat a následně například povolit čtení jediného adresáře:

```
User-agent: *
Disallow: /

User-agent: Googlebot
Allow: /web/
```

Tímto zápisem zakážeme všem robotům přístup kamkoliv na webu, ale povolíme Googlobotovi přístup do adresáře web. Problémem konstrukce Allow je, že není standardem, tudíž se nemůžeme spoléhat na to, že to budou respektovat všichni roboti. Avšak Google, Seznam a Jyxo Allow podporují. [10][1]

2.2.8 Kdy se objeví nová stránka ve výsledcích vyhledávání?

Google a ostatní stránky navštíví nové webové stránky automatické, je ovšem důležité, hlídat si, zda jsou dané stránky uvedeny někde na webu (například v katalogu) nebo je alespoň zaregistrovat přes

výše zmiňované registrační stránky jednotlivých vyhledávačů. Při umístění odkazů na jiné stránky, je dobré, přesvědčit se, že dané stránky vyhledávač zná.

Například Google po objevení nové stránky, stránku zařadí do databáze Everlux. Je to databáze nových stránek, kde se stránky uchovávají několik dnů (tři, čtyři nebo týden). Po uplynutí zmíněné krátké doby, stránka zmizí z výsledku vyhledávání a po několika dnech až týdnech (přibližně čtrnáct dnů) se stránka opět do výsledků vrátí. Tentokrát je stránka nahrána již v hlavním indexu. Po tomto přesunu již stránka není tak nahoře, jako předtím, zato je její umístění stabilnější.

Pro shrnutí. Nové stránky se po nalezení umísťují do vedlejší databáze, nikoliv do hlavního indexu. Z této databáze se nejčastěji dostanou do výsledku při hledání tzv. řídkých slov, to jsou slova a slovní spojení, která nejsou až tak běžná nebo se tolik nevyskytují. Po uplynutí nějaké doby, je web přesunut z vedlejší databáze do hlavního indexu, kde je už vyhledáván na běžná slova.

2.2.9 Jak sestavují vyhledávače popis webu ve výsledku?

Titulek odkazu, je sestaven z titulku na stránce, který se nachází mezi párovými tagy <title></title>, proto je důležité, aby byl title na stránce co nejvýstižnější a obsahoval klíčové slovo, na které chceme stránku optimalizovat. Zároveň by měl titulek být jedinečný, aby něčím přilákal, protože právě titulek často rozhoduje mezi tím, zda uživatel na odkaz, který mu nabídne vyhledávač ve výsledcích klikne nebo ne. [2]

Z čeho se vybere text v úryvku, není jednoznačně dáno. Každý vyhledávač si může jít svou vlastní cestou. Podle toho, jak který vyhledávač pracuje, zda si do databáze ukládá celou stránku, nebo jen klíčové slova, jdou vypsat možnosti, podle kterých vyhledávače nejčastěji vypisují text v úryvku:

- Zobrazují okolí klíčového slova. Je tedy zřejmé, že vyhledávače musí mít v databázi nejlépe obsah celého webu, aby toto okolí mohli vypsat. Obsah úryvku se většinou skládá z částí vět, které obsahují hledaný výraz, tento hledaný výraz, je pak zvýrazněn, nejčastěji tučným textem.
- Zobrazují obsah metaznaček. Vyhledávač zobrazuje obsah metaznačky Description (popis). Problémem tohoto zobrazení je, že do popisu si může tvůrce webu napsat cokoliv, i něco co neodpovídá obsahu stránky. Někdy se dokonce stává, že Description má větší obsah, než samotná stránka.
- Zobrazují popis z katalogů. Některé vyhledávače, jako například MSN, zobrazují popis z katalogu ODP, o kterém jsem se zmiňoval výše. Uživatel ovšem může tuto možnost zakázat pomocí metaznačky <meta name="robots" content="noodp" />, čímž zamezí vyhledávačům v zobrazení výsledku v popisu katalogu ODP. Jiné vyhledávače, obsahující vlastní katalogy, často používají popis z těchto vlastních katalogů.

2.2.10 Co ovlivňuje umístění stránek ve vyhledávání?

Vyhledávače určují výsledky zobrazení na stránce s výsledky vyhledávání podle mnoha kritérií, ale ne všechny kritéria jsou veřejnosti známé. Je zde zjevný důvod ochrany těchto kritérií, protože by se weby snažili optimalizovat pouze na tyto kritéria a vyhledávač by přestal být relevantním. Pro příklad uvedu několik základních, resp. známých kritérií:

- Vzájemná poloha nalezených slov (u víceslovných frází)
- Umístění nalezených slov (poloha v dokumentu)
- Umístění fráze nebo hledaného slova v titulu stránky, metaznačce Description, nadpisech
- Podle váhy stránky v očích vyhledávače
- Na základě počtů odkazů na stránky a kvality těchto odkazů

Negativní vlivy umístění:

- Stránky bez titulků
- Málo textu
- Žádné odchozí linky
- Příliš mnoho tučného textu
- Mnoho nadpisů (h1, h2, atd.)
- Duplikované stránky, tedy duplicitní obsah

Ještě bych rád objasnil zmíněnou polohu dokumentu, nejlepší bude vysvětlení na příkladu. Při hledání fráze „koupím auto“, kdy se slovo „koupím“ nachází na začátku stránky a „auto“ na konci, se pravděpodobně nejedná o hledanou stránku. Na druhou stranu, pokud se slova budou nacházet vedle sebe, je pravděpodobné, že se jedná o stránku, která je hledána.

Co může rozhodnout o tom, zda uživatel klikne na daný odkaz ve vyhledávači nebo ne, je fakt, že se například na prvních čtyřech místech objeví něco jiného, než co hledal. Co mu většinou na první pohled napoví, je právě titulek stránky. Proto je dobré, aby titulek stránky obsahoval opravdu přesnou a stručnou (50 – 80 slov) specifikaci stránky, pokud možno s klíčovým slovem dané stránky. Protože pokud například uživatel hledá „koupím BMW 3“ a první čtyři odkazy, budou mít titulek „Koupím BMW 3“, je jasné, že první odkaz, na který klikne, bude až pátý, který v titulku obsahuje „Prodám BMW 3“.

Podle Rand Fishina existuje 10 pozitivních a 5 negativních faktorů, které dokáží zásadně ovlivnit umístění stránky ve výsledcích vyhledávání:

+ Klíčové slovo použité v metaznačce title
+ Celková popularita stránky vyjádřená vysokým počtem stránek odkazujících na daný web
+ Text odkazů, které odkazují na váš web (anchor text)
+ Dobré hodnocení stránek uvnitř webu, nejen první stránka, ve spojení s dobrou strukturou
+ Starší stránky jsou považovány za relevantnější než nové
+ Obsahová relevance příchozích odkazů = odkazy z obsahově příbuznějších stránek jsou relevantnější

+ Popularita stránek v rámci obsahově příbuzné komunity
+ Klíčová slova na stránce
+ Hodnota odkazujících stránek – čím je odkazující stránka lépe hodnocena, tím je odkaz hodnotnější
+ Temo nárusti nových příchozích odkazů
− Stránky jsou často nepřístupné pro roboty a návštěvníky
− Obsah je velmi podobný nebo totožný s obsahem na jiných stránkách nebo doménách
− Příchozí odkazy nejsou valné hodnoty, přicházejí ze spamujících stránek
− Duplicitní metaznačky title a description na mnoha stránkách v rámci webu
− Zapojení do nejrůznějších link farem a systémů nákupu zpětných odkazů

Dalšími faktory, které můžou negativně ovlivnit umístění stránek je například změna algoritmů ve výpočtu výsledku, kdy jeden den je stránka na prvním místě ve vyhledávání na určitou frázi a za týden může být až na desátém místě. Tohle je jeden z faktorů, které jsou těžko ovlivnitelné. Nebo špatný hostingový server s velkou odezvou, nebo ještě hůř s velkou nedostupností v době, kdy zrovna roboti procházejí danou stránku.

2.2.11 Jak jsou dotazovány vyhledávače?

Důležitou informací na závěr je, jak se vlastně lidé ptají vyhledávačů na to, co potřebují najít. Důležité je to proto, že podle toho se může odvíjet optimalizace stránky na určitou frázi. Existují tři základní typy dotazování. [5]

1. **Navigační dotazování**

U navigačního dotazování, uživatel očekává, že výsledkem bude jedna konkrétní stránka. Pravděpodobně zná její adresu, momentálně adresu zapomněl, nebo jen předpokládá, že daná hledaná stránka existuje. Vyhledávače také zohledňují počet prokliků na jednotlivé odkazy. Při určitém hledaném výrazu je těžké, pokud je na danou stránku počet prokliků například 90%, takovouto adresu sesadit z prvního místa ve vyhledávání.

2. **Informační dotazování**

Uživatel hledá informace, které se nacházejí na několika webových stránkách, a není jasné, kterou přesně hledá. Většinou jsou to velice obecné výrazy, jako například auto, jídlo, dovolená, ptačí chřipka, apod.

3. **Transakční dotazování**

Jedná se většinou o dotazy, týkající se nákupu, stahování, telefonních seznamů, apod. Od uživatele se očekává, že na nalezené stránce bude dále provádět další činnost – transakci. Bohužel, co vyhledávače prozatím ještě nezvládly, je zvládnutí externích faktorů, které při tomto vyhledávání hrají podstatnou roli, jako je například cena produktu.

2.3 Optimalizace zdrojového kódu stránek

V této kapitole se budu soustředit na výše zmiňované tzv. on-page faktory stránky, tedy to, jak stránky kvalitně napsat, aby se vyhledávačům líbily a byly umístěny pokud možno co nejvýše ve výsledcích vyhledávání.

2.3.1 10 zásad při tvorbě stránek

1. Validní HTML kód (http://www.validator.w3.cz) spolu s funkčními odkazy na stránce. Při výskytu nefunkčních odkazů, nemusejí být roboti schopni správně indexovat stránky.
2. Při přesunu stránek, nastavit starou URL adresu tak, aby uživatele (robota) přesměrovala na novou URL adresu.
3. Jistota, zda roboti nemají zakázaný přístup na stránky (viz kapitola: Spolupráce a komunikace s pavouky)
4. Používání statických URL adres. Generované, složité a často se měnící adresy nemají roboti, stejně jako uživatelé, rádi.
5. Kvalitní a hodnotný obsah.
6. Do viditelného textu stránky, zahrnout klíčová slova, která mohou uživatele vyhledávat při vyhledávání ve vyhledávačích.
7. Rozumná velikost stránek. Stránka bez obrázků by neměla přesáhnout velikost 100kb, ideální velikostí je, pokud stránka bez obrázků bude menší než 40kb.
8. Každá stránky, by měla být dostupná, alespoň z jednoho statického odkazu.
9. Text, který má být indexován, musí být umístěn mimo obrázky. Roboti neumějí číst obrázky, takže při například názvu firmu, umístěném pouze v logu, tedy obrázku, nebude robot vědět, jaká firma se na stránce nachází.
10. Jednoduchá hierarchie webu. Všechny stránky by měly být dostupné z domovské stránky na maximálně tří prokliky, ideálně jeden až dva prokliky. Dobré je, přidat mapu webu, aby se roboti dostali i na stránky, které nejsou lehce dostupné.

2.3.2 Klíčové slova a metaznačky

Relevantní klíčové slova

Důležitou myšlenkou, při volbě, na které klíčové slovo optimalizovat stránku, je, abychom neoptimalizovali stránku na slovo, které je nejčastěji vyhledávané ve vyhledávačích, ale aby byla stránka optimalizována na slovo, které ji nejvíce vystihuje. Například, pokud máme stránku, která patří autobazaru v Ostravě, je zbytečné, možná až nežádoucí, optimalizovat stránku na slovo „autobazar“. Při takovéto optimalizaci, bude zaprvé problematické, dostat se na přední příčky ve vyhledávačích, tudíž i drahé. Zadruhé, i když bude tato fráze často vyhledávána, tak stránce nepřinese žádný užitek, protože, pokud bude frázi autobazar vyhledávat například člověk z Prahy, těžko ho potěší, že nalezl autobazar v Ostravě. Takovýto uživatel pravděpodobně stránku opustí, jakmile zjistí, že se jedná o Ostravu, což firmě zisk nepřinese, naopak je pravděpodobně poškozená, při investici na optimalizaci na tak náročné klíčové slovo. Ovšem pokud zvolíme optimalizaci na klíčové slovo

„autobazar Ostrava“, nebude optimalizace až tak náročná, tudíž levnější a zároveň firmě přinese větší užitek, v podobě zákazníků, kteří vyhledávají stránky autobazaru v Ostravě, i přes skutečnost, že tuto frázi nebude vyhledávat tolik lidi, jako frázi „autobazar“.

Výstižná charakteristika z knihy Michala Kubíčka [1]:

„Klíčem úspěchu je nalezené správné rovnováhy mezi přesností a popularitou“.

Strategie long tail (dlouhý ocas)

Marketingoví odborníci tvrdí, že zákazníci jsou jako kometa, která není tvořena jen hlavním tělem komety, ale i velkým počtem malých segmentů, dlouhým ocasem, tzv. long tail. Z toho plnye, že množství potencionálních zákazníků, kteří se nacházejí v ocasu komety je mnohem větší, než množství uživatelů, které je schopna stránka oslovit v hlavním těle komety.

Tuto strategii často využívají především internetové obchody. Spoléhají na to, že díky tomu, že nemusí držet velké sklady se zbožím, tak mohou mít větší nabídku, tudíž mohou uspokojit více zákazníků i při prodeji menšího množství jednotlivých kusů zboží.

Při tvorbě klíčových slov, je také důležité vědět, že při různé fázi nákupu uživatel vyhledává na jiné klíčové slova. S přibývajícími informacemi uživatel konkretizuje vyhledávanou frázi. Například při nákupu fotoaparátu, bude počáteční obecná fráze uživatele „digitální fotoaparát“. Po zjištění malého množství nových informací uživatel přejde na frázi „širokoúhlý digitální fotoaparát“. Nakonec do fráze zahrne třeba i značku, možná i typ.

Výběr klíčových slov

Kvalitním nástrojem, který nám pomůže při výběru, na jaké klíčové slovo optimalizovat, je nástroj od Googlu na stránkách *https://adwords.google.com/select/KeywordToolExternal*, který nám pomůže při výběru a rozhodování na jaké klíčové slovo stránku optimalizovat. Zadáme si zde o jakou frázi se zajímáme a program nám nabídne podobné fráze s informacemi, kolik lidí za den a měsíc tuto frázi vyhledává a jaká je konkurence v dané frázi.

Nástroj je užitečný, nejen v těchto případech:

- Hledání nových klíčových slov a podobných frází
- Typické vzorce, jak se uživatelé konkrétně ptají vyhledávačů
- Možné doplňky, pro aktuální klíčová slova
- Zjištění, jak frekventovaně jsou fráze vyhledávána
- Nové spojení, pro obsahově vyčerpané stránky

Výběr na jaké klíčové slovo optimalizovat je tou nejdůležitější součástí, protože pomáhá uživatelům internetu, nalézt právě onu konkrétní stránku. Pro výběr klíčových slov, lze také použít nástroj *SEO Administrator*. V něm použít nástroj, zvaný *Keyword suggestion*. Program má přístup ke zdrojům (databázím) různých aplikací, jako jsou například Overture nebo Wordtracker. Program vám pomůže s analýzou konkurenčních stránek a tím vám napoví, na jaké klíčové slova, by se měla stránka zaměřit.

Svů vlastní nástroj má také vyhledávač Seznam. Tento nástroj spustil v únoru roku 2008 a zařadil jej do svého reklamního systému Sklik.

Za zmínku také stojí český systém PPC Etarget, nacházející se na adrese *http://www.etarget.cz/ customer/info/stats.php ?cmb=1*. Program zobrazí fráze, které jsou vyhledávány na onternetu, pro vámi zadané slovo. Ukazuje nejenom tvary frází, ale také počet vyhledávání za den a počet inzertů. Pokud je počet inzertů větší než jedna, můžeme si zobrazit, po kliknutí na danou frázi, které stránky jsou na tuto frázi zaměřené. Nástroj je lokalizován na Česko, Slovensko, Maďarsko, Rumunsko, Srbsko, Bulharsko a Chorvatsko.

V neposlední řadě ještě zmíním nástroj od portálu Centrum.cz s adresou *http://www.adfox.cz/ pomocnikpronavrhkw.phtml*. tento nástroj rozlišuj velká a malá písmena a slova s a bez diakritiky. Je užitečný nejen proto, že zobrazuje široký záběr synonym k zadanému slovu.

Tematická analýza

Tematická analýza se dělí na tzv. *vertikální* a *laterální*. Vertikální analýza je proto, že jde napříč oborem a je to rozbor klíčových slov, tematicky související se stránku. Pro lepší pochopení, vysvětlím na příkladu, například internetový obchod prodávající potřeby pro bojové umění. S touto stránkou budou tematicky ladit slova, týkající se jak potřeb, tedy kimono, rukavice, atd., tak samotné názvy bojových umění a jejich slovních spojení, jako judo, kimono na judo atd.

Oproti tomu, laterální analýza vyhledává slova, která jsou příbuzná s oborem podnikání dané stránky, ale nejsou přímo spjatá s tímto oborem. U výše zmíněného internetového obchodu s potřebami pro bojové umění, může například potencionální zákazník hledat slovní spojení, jako kurzy sebeobrany, apod. Díky takovýmto slovním spojením, můžeme pomocí výměny zpětných odkazů s jinými stránkami, na zmíněné stránky dostat potencionální zákazníky. Dalšími slovy, můžou být různá slovní spojení, jako hadžime, šijak, apod., nebo informace o bojových uměních, které může potencionální zákazník vyhledávat.

Řeč zákazníků

Důležitým aspektem při výběru slov je, aby slovo odpovídalo řeči zákazníka. Například taková lednička, je spisovně chladnička, ale kolik lidí na internetu bude vyhledávat výraz chladnička namísto lednička? Takže při výběru relevantních klíčových slov, jsou důležité tyto body:

- Jsou vybraná slova obecně užívaná?
- Jsou to slova, na které přijde zákazník? (ne pouze návštěvník, ale ten, kdo s námi uzavře obchod)
- Jaké klíčové slova používají úspěšní konkurenti?
- Jaká je frekvence vyhledávání u jednotlivých slov?

Jaké fráze lidé zadávají do vyhledávačů?

Dalším důležitým aspektem, při výběru na které klíčové slovo nebo klíčovou frázi bude stránka optimalizována, je fakt, jak se lidé vyhledávačů ptají jednoslovně, dvouslovně či víceslovně? Na toto téma v roce 2006 zveřejnila společnost OneStat.com výsledky dlouhodobého průzkumu, který říká:

- 28,91% lidí se ptá vyhledávačů dvouslovně
- 27,85% lidí se ptá vyhledávačů tříslovně
- 17,11% lidí se ptá vyhledávačů čtyřslovně
- 11,43% lidí se ptá vyhledávačů jednoslovně
- 8,25% lidí se ptá vyhledávačů pětislovně
- 3,68% lidí se ptá vyhledávačů šestislovně
- 1,59% lidí se ptá vyhledávačů sedmislovně

[14]

Ze statistik také vyplývá, že většina lidí se neobtěžuje s velkými písmeny. Ostatně tuto skutečnost nerozlišují ani vyhledávače. Stejně je na tom i jednotné a množné číslo a různé druhy skloňování slov. Pro tyto slovní převody, jsou ve vyhledávačích sestaveny speciální slovníky, za pomocí odborníku na český jazyk.

Co se týče diakritiky, tak se výsledky ve vyhledávačích mírně liší, je to především dáno tím, že v českém jazyce, může mnohdy nastat situace, kdy slovo s diakritikou má jiný význam než slovo bez diakritiky.

Kde zjistit frekvenci vyhledávání klíčových slov

Pro zjištění frekvence vyhledávání klíčových slov, je nejlépe použít kombinaci různých nástrojů, jako jsou Našeptávač od Seznamu, dále systémy statistik PPC Google Adwords, AdFox, Sklik a ETARGET.

Našeptávač od Seznamu

O této funkci seznamu, se vedou na dlouhé diskuze, zda je objektivní nebo ne. Rozhodně se tato funkce nemůže považovat jako klíčová, ale pro orientaci na internetu to není špatná pomůcka. Sami tvůrci tohoto programu přiznávají, že některé nevhodné nebo vulgární slova, jako například *sex* upravují, nebo přímo z našeptávače odstraňují.

Kde tuto aplikaci hledat? Každý určitě ví o aplikaci, kdy Seznam, při vyhledávání jistého slova, nabízí pod vyhledávacím políčkem slova nebo fráze jemu podobné. Přesně tohle je aplikace Našeptávač. Před rokem 2007 se u těchto frází také vyskytovalo číslo, které znázorňovalo počet hledání dané fráze za den. V dnešních dnech, již tuto statistiku na svém místě nenalezneme, ale můžeme ji nalézt v aplikaci *Seznam Lištička*, kterou je možné stáhnout a nainstalovat z adresy *http://www.listicka.cz*. Tato aplikace neukazuje přesné hodnoty hledání za den, ale spíše orientační hodnoty, zprůměrovaný za přesně nespecifikované časové období.

Pro zobrazení výsledků hledání za den, slouží aplikace, která je doplňkem pro prohlížeč *Mozilla Firefox* od autora Marka Prokopa. Tato aplikace je k stažení a nainstalování na adrese

http://userstyles.org/styles/2578. Pomocí této aplikace, se nám opět objevuje, jako před rokem 2007, na doméně *http://search.seznam.cz* počet vyhledávání zadané fráze za den.

Výhodou aplikace našeptávač je, že nenabízí tvůrcům a optimalizátorům webových stránek, jen statistiky počtu vyhledávání určitých frází za den, ale také to, co je uživatelům předkládáno v nabídce, při vyhledávání. Tímto, se tento nástroj pro vývojáře stává velice silným nástrojem, pro výběr a optimalizace klíčových slov na stránce. Ovšem na druhou stranu, je tato možnost také velkou slabinou Seznamu, protože je relativně lehce zmanipulovatelná. Při velkém množství dotazovaní na určitou frázi, je daná fráze uživatelům Seznamu nabízena, jako možnost v Našeptávači. [4]

Google Suggest

Stejně jako u Seznamu se jedná o našeptávač, ale od vyhledávače Google. Je třeba podotknout, že s našeptávačem nepřišel první Seznam, ale samozřejmě Google, seznam se od Googlu, ostatně jako v mnoha věcech, inspiroval a jeho myšlenku převzal. Je to silný nástroj, který mnohdy ovlivňuje to, co uživatelé nakonec vyhledávají. Bohužel, stejně jako Našeptávač od Seznamu, i Google Suggest skryl počet hledání nabídnuté fráze, před uživateli. Základním předpokladem k tomu, aby se dostala určitá fráze do nabídky Google Suggest je samozřejmě počet hledání. Ovšem není to jediným kritériem dokonce neovlivňuje ani pořadí nabídnutých frází.

Řazení

Jak jsem již zmínil, frekvence vyhledávání určité fráze, není klíčem k seřazení v Google Suggest. Klíčem k tomuto seznamu je popularita, které je ovšem zahalena tajemstvím.

Mimo popularitu ještě rozhoduje tzv. *freshness layer*. Jedná se o ovlivnění našeptávače aktuálními událostmi, pokud se například náhle změní počet vyhledávání u určité fráze, je tato fráze zařazena výše v nabídce frází. Tuto skutečnost pravděpodobně ovlivňuje také *Google news*, kdy při nějaké velké události, našeptávač usnadní uživatelům vyhledávání. Dalšími ovlivňujícím faktorem je *Personalized searches*, který ovlivňuje vyhledávání, pokud je uživatel přihlášen ke službě Google. Poté jsou výsledky ovlivňovány kamarády, zájmy, apod.

Opravování pravopisu

Další funkcí Google Suggest je, že nám v nabídnutých frázích, nabízí pravopisně opravené výsledky. Například při zadání slova „austobazar“ nám našeptávač automaticky nabídne slovní spojení s opravěným slovem „autobazar“, „autobazar ostrava“, atd.

Geografické cílení

Z předchozího příkladu u opravy pravopisu vyplývá také skutečnost, že se slova v nabídce liší také podle toho, ve kterém městě, nebo kraji danou frázi vyhledáváme. Například v Ostravě na slovo „autopůjčovna“ našeptávač nabídne „autopůjčovna ostrava“, ale v Českých Budějovicích je v nabídce „autopůjčovna české budějovice“.

Filtrování výsledků

Google si také dává pozor na to, aby se, jak se tomu v minulosti stalo, díky nabídnutým frázím nedostal do konfliktu se zákonem.

Ochrana před nesnášenlivostí a rasovou nenávistí

Co je dalším filtrem, je ochrana základních lidských práv a svobod. Takže při napsaní slova „nesnáším“ Google Suggest nenabídne žádné etnické skupiny, ale například „loučení, školu, když, atd.“. Podobně je tomu i u slova „nemám rád“ a další podobné slova. Mezi výrazy, které Google u nesnášenlivosti nenapoví, patří:

- Rasy a etnikum
- Barvy
- Národnosti
- Náboženského vyznání
- Postižené
- Gender a genderovou orientaci
- Lidi různého věku
- Veterány

Soudně vymahatelné škody

Google blokuje některé fráze, protože často byl souzen za nevhodné našeptávání. Například u slov „podvod“, „podvodník“ může být výsledek zavádějící anebo přímo urážející. Proto si u takovýchto slov, Google dává pozor, aby za nimi nenásledovaly jména osob, nebo názvy firem. Existují ovšem výjimky, jako například výraz „podvody na aukru“, kdy bude mít Google pravděpodobně s firmou smlouvu, nebo jsou vymezeny některé hranice, co má Google blokovat a co ne.

Kontroverzní výrazy

Dalším ovlivňujícím faktorem, jsou kontroverzní výrazy jako „rasismus“, “křesťanství“, „islám“, apod. U těchto výrazů je našeptávač očištěn od nevhodných slov.

Ochrana před nelegálním stahováním a filmy pro dospělé

Google se také snaží o zablokování některých známých serverů, nebo warezfór pro sdílení nelegálního obsahu na internetu. Takže například při zadání názvu fóra, nám nepředloží nabídku s názvy filmů. Podobně je tomu i u filmů pro dospělé.

Statistiky vyhledávaných slov ze systémů PPC

Jako nejpřesnější statistiky, se považují právě statistiky ze systémů PPC. Ovšem nic není zadarmo a t se týká i systémů PPC. Na druhou stranu, peníze, které uživatel vloží do PPC, mu přinesou užitek v podobě nových zákazníků (PPC – pay per click, neboli platba za proklik).

Systémy PPC, ze kterých se dá zjistit statistika vyhledávaných slov:

- Google Adwords – zobrazuje se ve výsledcích vyhledávání Google
- Sklik – zobrazuje se ve výsledcích vyhledávání na Seznamu

Systémy PPC, které se zobrazují ve vyhledávání a na partnerských stránkách (statistiky vyhledávaných slov mohou být zkreslené):

- Adfox – systém PPC portálu Centrum
- Google AdSense

Systémy PPC, zobrazující se pouze na partnerských stránkách (nejsou použitelné, pro statistiky vyhledávaných slov):

- ETARGET – zobrazuje se na českých serverech, jako je například iDNES
- BBtext – provozovatelem je největší český výměnný systém Billboard

Žebříčky nejhledanějších slov

Na internetu je mnoho různých žebříčků nejhledanějších slov, za daný časový úsek. Vyhledávače jako jsou Seznam a Google, však tyto výsledky velice nerady zveřejňují. Výhodou, ale i nevýhodou systémů zveřejňující žebříčky nejhledanějších slov je, že jsou to opravdu nejhledanější slova a povětšinou již plně zabraná velkým konkurenčním bojem o umístění ve vyhledávačích.

ETARGET

Seznam 100 nejhledanějších slov, na českém internetu, za posledních 7 dní. Tyto statistiky nalezneme na adrese *http://www.etarget.cz/customer/info/stats.php?&t100=1*.

Seznam

Seznam již zrušil stránku, která se nacházela na stránce *http://katalog.seznam.cz/top_keyword.html* a která zobrazovala žebříček nejhledanějších slov, jejich pokles a nárůst a současně i statistiku hledání za den. Zároveň jak je výše zmíněno, tak Seznam zrušil statistiku vyhledávání slova v našeptávači. Díky tomu, je momentálně nejvyužívanější možností na seznamu, nenápadný odkaz „Statistika hledaného výrazu“, který se nachází vždy v patičce dokumentu, zobrazující výsledky hledání u hledaného výrazu, nebo na adrese *http://search.seznam.cz/stats?collocation=klicove-slovo* (klicove-slovo nahradíme námi zvoleným klíčovým slovem). V této statistice je zobrazen:

- Statistika hledanosti výrazu
- Rozšířená shoda hledání (dotazy obsahující dané klíčové slovo)
- Přesná shoda (dotazy v přesném znění)
- Nejhledanější dotazy obsahující dané klíčové slovo

Rozdíl mezi našeptávačem a touto aplikací je, že našeptávač sleduje pouze přesnou shodu určitého výrazu. Naproti tomu, tato aplikace sleduje i slova s daným klíčovým slovem spjatá, nebo příbuzná.

AdFox (Centrum.cz)

Statistiky, zveřejňující portál Centrum se nachází na adrese *https://www.adfox.cz/nejhledanejsislova.phtml*. Další statistiky tohoto portálu se nachází na adrese *http://www.adfox.cz/statistiky.phtml*.

Umístění klíčových slov

Z předcházejících kapitol, jsme schopni, zjistit si statistiky a konkurenci u jednotlivých klíčových slov. Nyní musíme klíčové slova zapracovat do naší stránky. Kam tedy klíčové slova umístit? Základním pravidlem, od autora Michala Kubíčka je, cituji:

„klíčová slova se na vaší stránce musí vyskytovat“.

I když tato citace může znít směšně, někdy se stává, že na stránkách se klíčové slovo sice vyskytuje, ale třeba jen v podobě obrázku nebo flash animace.

Řešením obrázků na stránce jsou alternativní popisy obrázku. Je to text, vyskytující se v atributu alt u elementu img a má za úkol, vyhledávačům objasnit, co se nachází na daném obrázku nebo co obrázek znázorňuje.

Ještě kvalitnějším se obrazek stane, pokud obrázek není označen jen nic neříkajícím kódovým označením, ale přímo tím, čím je, to se týká i odkazu u v elementu a.

Některé stránky řeší nedostatek textu na stránkách „nekalým řešením“ a to tak, že umístí na stránky hodně textu, ale text má stejnou barvu jako pozadí stránky a je psán co nejmenším písmem. Není to zrovna optimální řešení, protože dnešní vyhledávače si toho umí všimnout a poté se stránka dostane spíše do problémů v podobě zabanování nebo zařazení do blacklistu.

<body bgcolor=“blue“>mnoho opakujícího se textu, tedy spamu, obsahující klíčové slova

Kde vyhledávače klíčové slova vidí:

- V URL adrese (*www.klicove-slovo.cz*, *www.domena.cz/klicove-slovo*)
- V hlavičce stránky
    - Titulek stránky <head><title>Klíčové slovo</title></head>
    - Obsahové metaznačky jako <description>, <keywords>
- V těle stránky
    - Alternativní popisek obrázků, tedy atribut alt v elementu img
    - Nadpisy (<h1>, <h2>, <h3>, <h4>, <h5>, <h6>)
    - Odstavce (<p>)
    - Titulky odkazů (<a title=“kličové slovo“>)
    - Do textu odkazů, tzv. anchor text (Klíčové slovo)
    - Do zvýrazněného textu, tedy , , <b>, <i> (vyznám elementu strong je vyšší než b)
    - Do seznamů (<ul>, <li>, <dd>, <dt>)

Kde vyhledávače slova ignorují:

- V obrázcích
- V elementech vytvořených JavaScriptem anebo AJAXem
- V animacích, obrázcích a dalších prvcích vytvořených flashem

Výše zmíněný seznam, nelze brát jako standart, protože každý vyhledávač má jinačí možnosti a pohlíží na stránku trochu odlišně. Například takový Google tvrdí, že flashové aplikace parsovat umí, nicméně podotýká, aby byl přiložen i textový soubor.

Danny Sullivan, šéfredaktor Search Engine Watch:

> *„Webové stránky ve Flashi působí stejně, jako kdybyste se představovali nepopsanou vizitkou.“*

Titulek stránky

Christine Churchill:

> *„Pokud byste měli čas jen na jedno SEO opatření na svém webu, věnujte ho tvorbě dobrých titulků stránek.“*

Titulek stránky patří mezi nejdůležitější prvky na stránce. Nejen pro to, že jej vyhledávače zobrazují v odkazu na danou stránku. Dobře zvolený titulek stránky, který vystihuje, tedy koresponduje, s obsahem stránky je nejsilnějším a zároveň nejlépe ovladatelným nástrojem stránky.

Umístění klíčového slova pouze v titulku a adrese stránky je však nedostačující. Při takovéto aplikaci, je stránce snížena její relevantnost a poté se moc vysoko ve výsledcích nezobrazuje. Nevhodným řešením také je, když je na celém webu jen jeden titulek stránky, který se při přechodu mezi podstránkami nijak neliší. Titulek by měl vždy vystihovat stručnou charakteristiku stránky, na které se nachází. [1]

10 zásad pro klíčová slova

1. Klíčové slovo v URL
    - některé vyhledávače přikládají názvu domény větší váhu, než klíčovému slovu obsaženému v URL adrese
2. Klíčové slovo v doméně celkem
    - rozdělovat jednotlivá slova pomlčkami (*http://www.hodinovy-manzel-olomouc.eu*)
3. Klíčové slovo v elementu title
    - Optimální délka je 10-70 znaků
4. Klíčové slovo v meta-elementu description
    - Délka textu do 200 znaků
5. Klíčové slovo v meta-elementu keywords
    - Ne více než 10 slov
    - Každé slovo, použité v tomto elementu se musí vyskytovat v těle stránky, jinak hrozí penalizace za irelevanci
    - Žádné slovo by nemělo být použito více než jednou, opět hrozí penalizace
    - Google tomuto elementu nepřikládá žádnou váhu, ale například Centrum ano
6. Frekvence klíčových slov v těle stránky by se měla pohybovat mezi 5-10% (poměr všech klíčových slov ke všem slovům na stránce)
    - Přílišná koncentrace klíčových slov vede k domněnce, že se jedná o spam
    - Individuální frekvence jednotlivých klíčových slov pak je 1-6% (poměr klíčového slova ke všem slovům)

7. Klíčové slova v nadpisech <h1>, <h2> a <h3>
8. Klíčová slova zvýrazněná pomoci a <b>
9. Blízkost klíčových slov v textu (pro klíčové slovo „levné kuchyně“ je lepší text „levné kuchyně a kuchyňské linky“ nežli „levné kuchyňské linky a kuchyně“)
10. Pořadí klíčových slov

2.3.3 Názvy domén a URL adres

Doména

Název domény by měl souhlasit se zaměřením webu stejně tak, jako je tomu například u knih. Názvy domén by měly být přinejlepším krátké, výstižné a lehce zapamatovatelné. Ne vždy, však můžeme takovéhoto názvu domény dosáhnout, jelikož jsou názvy jedinečné a v dnešní době jich již moc volných není (myšleno v oboru podnikání daného webu).

Marek Prokop, H1:

> *„Z pohledu SEO mají klíčová slova v doméně v podstatě tři významy: Jako text zpětného odkazu, nejčastěji na úvodní stránku, ale i na podstránky – pak je důležitá kombinovatelnost slov v doméně s dalšími slovy do konkrétnějších frází. Totéž ale na úrovni klíčových slov v URL. A nakonec pro navigační dotazy - pak je důležitá přesná shoda dotazu (případně s vynechanými mezerami) a doménového názvu.“*

Například při webu, na kterém budou obsaženy informace o firmě zabývající se odvětvím zvaným Hodinový manžel, budou přijatelné domény typu: *www.hodinovymanzel.cz*, *www.hodinovy-manzel.cz*, *www.manzel-na-hodinu.cz* apod.

Nicméně, naprostá většina vyhledávačů tvrdí, že jim na názvu domény nezáleží. Otázkou tedy zůstává, proč volit název domény, který obsahuje klíčové slova? Jednoduše proto, že se často u zpětných odkazů vyskytuje právě název domény, tedy web, který na vás odkazuje, nebude odkazovat anchor textem „Hodinový manžel“, ale pravděpodobně textem „www.hodinovy-manzel.cz“. [4]

Pomlčky v doméně

Nejlepší odpovědí v otázce, zda registrovat doménu s pomlčkou či bez je, zaregistrovat obě domény. Je to hlavně kvůli dvěma faktorům. Prvním ovlivňujícím faktorem je, že doména s pomlčkou se sice v reklamě a na plakátech propaguje lépe, ale mnoho uživatelů dodnes neví, jak se vlastně pomlčka na klávesnici píše. Druhým ovlivňujícím faktorem je skutečnost nekalých praktik a to takových, kdy si jedna firma zaregistruje adresu *www.hodinovymanzel.cz* a druhá, konkurenční firma si zaregistruje adresu *www.hodinov-manzel.cz*, jenom proto, aby využívala překliků a mohla tak získat potencionální zákazníky první firmy.

Pomlčky jsou v názvu domény chápány jako oddělovače slov. Ale vyhledávače se k názvům domén s pomlčkou a bez staví tak, že při vyhledávání fráze, z této fráze odstraní mezery, tedy vytvoří jedno dlouhé slovo spojené z více slov, které poté porovnávají s názvy domén. Takže při vyhledávání výrazu „hodinový manžel“ nám vyhledávač zobrazí jak adresu „*www.hodinovy-manzel.cz*“ tak „*www.hodinovymanzel.cz*“.

Musím tlumočit názor mnoha odborníků, že klíčové slovo v doméně je důležité, ale nic se nesmí přehánět. Proto není zrovna vhodné, zvolit za název domény pro prodej zahradnických prostředků „*www.zahradnictvi-prodej-sekacky-nuzky-ploty-sazenice.cz*“.

Jonah Stein: „*Jestliže vaše dostatečné stará, bez pomlčková doména obsahuje vaše primární klíčové slovo, čtvrtinu cesty do Top 10 máte za sebou.*“

Řešením stávající situace, kdy je nedostatek volných a zároveň vhodných názvu domén může být subdoména. Například při prodeji bazénu, si uživatel již nezaregistruje doménu *www.prodej-bazenu.cz*, ale může si zaregistrovat doménu *www.bazenu.cz* a pomocí subdomén vytvořit adresu *prodej.bazenu.cz*, *montaz.bazenu.cz*, atd.

Doména v řeči zákazníků

Existují domény typu *www.web4u.cz*, kde ne každý potencionální zákazník ví, že „4“ znamená „for“ a „u“ znamená „you“. Podobné můžou být i například cizojazyčné firmy jako je Wüstenrot, kdy pro německy hovořící uživatele je přirozené, napsat do adresního řádku *www.wuestenrot.cz*, ale většina německy nehovořících zákazníků napíše *www.wustenrot.cz*.

Stáří domény

Stáří domény, je také důležitým faktorem při určování relevantnosti výsledku. Čím starší doména je, tím je považována za relevantnější. Stáří domény není dáno dnem zaregistrování domény, ale prvním zaindexováním vyhledávačem, zjištěním zpětných odkazů a odpovídáním na určitý dotaz ve vyhledávání.

Koncovka domény

Ideální koncovkou domény v naši geografické poloze je .cz a to proto, že některé vyhledávače určují primárně jazyk domény podle koncovky. Donedávna tomu tak bylo i u Seznamu, avšak nyní prohledává i jiné koncovky. Pokud je tedy doména umístěna například na stránce s koncovkou .com, je třeba zažádat Seznam o indexaci takovéto česky psané stránky. Co se týče rozdílu, tak vyhledávače nerozlišují mezi koncovkami. Google indexuje bez rozdílu všechny domény.

URL adresa

SEO friendly URL

SEO friendly URL, nebo také *cool URI*, je jakýsi standart, jak by měla vypadat vhodná URL adresa. Nejlépe vystihne tyto pravidla vysvětlení na příkladu. Máme adresu *http://obchod.cz/shop_cat.php?sekce=2&id=198*, tuto adresu bude lépe převést na *http://obchod.cz/skoda/tlumice*. Pro vyhledávač, ale i pro uživatele, je tedy lepší, když se adresa tváří jako statická, nežli dynamická. Pro uživatele je tato adresa mnohem lépe zapamatovatelná.

Dalším důležitým aspektem je, aby se URL adresa neměnila. Pokud je to nezbytné a URL adresa se musí změnit, například v důsledku přejmenování článku, měla by stará adresa přesměrovávat uživatele a roboty na novou URL adresu.

Diakritika v URL

U URL s diakritikou nastává problém, protože URL nemá přesně definovaný *charset*, tedy znakovou sadu. Takže při použití diakritiky, jsou písmena s diakritikou nahrazovány speciálními znaky a stránka tak ztrácí na své přehlednosti a zapamatovatelnosti. Lepším řešením, je stránky obsahující diakritiku překládat na stránky bez diakritiky, kdy písmeno například písmeno „č“ nahradíme písmenem „c“ a ne zástupným znakem „%C4%8D“, jak je tomu v kódování UTF-8.

Adresa s WWW nebo bez?

U hostingových serverů bývá pravidlem, že pokud správce webové stránky neučiní jinak, tak bere stránku s WWW i bez na začátku za totožnou. Takže při zadání adresy s WWW nebo bez, přesměruje uživatele na adresu, na které se nachází obsah. Namísto toho vyhledávače chápou adresy s WWW a bez za dvě rozdílné adresy. Je tomu tak proto, že WWW na začátku adresy, je chápána jako subdoména dané domény.

Z pohledu SEO se vyplatí, soustředit se pouze na jednu adresu. Ale pokud se tedy správce soustředí na adresu *http://hodinovy-manzel.cz/* tak asi nezabrání tomu, aby na něj někdo na webu odkazoval pomocí adresy *http://www.hodinovy-manzel.cz/*. Vyhledávač při naražení na takový to odkaz projde stránku, která jeje přesměruje na vaši stránku neobsahující WWW, zaindexuje tuto stránku jako stránku obsahující WWW a uloží do databáze. V databázi tak bude mít dva rozdílné weby, ale se stejným obsahem. Díky tomu, že se vyhledávače brání zahlcování své databáze duplicitním obsahem, tak při pravidelné kontrole zjistí, že se jeden web v jeho databázi nachází dvakrát. Zvolí jeden web a ten ze své databáze odstraní. Bohužel správce tohoto webu, nemůže ovlivnit, který web z databáze vyhledávače bude odstraněn.

Z tohoto faktu plyne jedna pro správce a optimalizátory nešťastná skutečnost. Vezměme si příklad, kdy na daný web odkazuje 200 zpětných odkazů z jiných stránek, kdy je polovina s WWW na začátku adresy a druhá polovina bez. Vyhledávač zaindexuje obě stránky jako různé a po zjištění duplicity jednu smaže. V této chvíli správce daného webu přišel o 100 zpětných odkazů. Podobně je tomu i u vnitřních odkazů, kdy jeden odkazuje na stránku *http://www.domena.cz/* a druhý na tutéž stránku, ale odkazem *http://www.domena.cz/index.html*. Vyhledávač opět tyto odkazy chápe jako rozdílné a po zjištění duplicity jeden smaže.

Přesměrování adres a domén

Při změně ať už celé domény, nebo adresy, je důležité, aby stará adres automaticky přesměrovávala lidi a roboty na novou adresu. Důležité je to proto, že po změně adresy, bude vyhledávač ještě relativně dlouho dobu ve svých výsledcích nabízet starou adresu. Poté se může stát, že potencionální zákazník přijde na adresu, ke jej přivítá pouze stránka s chybovým hlášením o neexistenci dané stránky.

Stránka 404

Pokud robot narazí na stránku s odpovědí 404, tedy neexistující stránku, úplně ji nezavrhne, ale ještě několikrát se na danou stránku vrátí, aby zjistil, zda se nejedná pouze o dočasný výpadek. Pokud chce správce robotovi dát vědět, že na stránku již nemusí chodit, tedy byla zrušena, musí mu odpovědět kódem 410.

Přesměrování na adresu s WWW

Pokud správce chce, aby všechny odkazy z internetu směřovaly uživatele a roboty pouze na adresu s variantou s WWW na začátku, a to nejenom dočasně ale trvale (musí odpovídat kódem 301), může tak učinit pomocí souboru *.htaccess* nebo *httpd.conf*. Provede to pomocí následujících dvou příkazů:

```
RewriteCond %{HTTP_ HOST} ^vase-domena.cz
RewriteRule (.*) http://www.vase-domena.cz/$l [R=301, QSA, L]
```

Tento zápis musí být v kořenovém adresáři domény, pokud by tomu tak nebylo, proměnná $1, by vyhodnocovala relativně k adresáři.

3. SEO Analyzátor

Další část této práce se zabývá informačním systémem, který slouží k analýze stránek z pohledu SEO. Informační systém stahuje weby, parsuje je podle výše zmiňovaných pro SEO významných elementů a jejich atributů. Z takto rozparsovaného textu vytváří klíčové slova za pomocí techniky n-gramů, klíčovým slovům je nastavena důležitost dle toho, v jakém elementu nebo atributu se vyskytují. Nad získanými hodnotami je provedena analýza, která nám zobrazí podobnost dvou a více stránek na u konkrétního klíčového slova, celé stránky a je zde i možnost porovnání podobnosti dvou domén.

3.1 Návrh přizpůsobitelného IS

IS je navržen tak, aby si každý uživatel mohl individuálně nastavit, které elementy a atributy jsou pro něj důležité a které nikoliv, zda se uživatel bude řídit pravidly pro určité roboty, jako například Googlebot, nebo zda tyto pravidla bude ignorovat, s tím, že univerzální pravidla, která mají zablokovat přístup všem robotům, jsou aplikací tolerována vždy. Poté si navolí, jakou velikost n-gramu chce v dané analýze provádět, zda respektovat češtinu, seřazení slov, apod. Dále je zde na výběr, z několika možností výpočtu a nad shromážděnými daty.

3.2 Funkční požadavky systému

3.2.1 Základní funkce IS

Procházení a parsování webů. Aplikace prochází všechny vnitřní odkazy, zadané webové stránky. Vytvoří si seznam všech vnitřních url adres stránky, které potom prochází a parsuje. Po rozparsování vytvoří z takto získaného textu klíčové slova (n-gramy), které následně převáží podle toho, ve kterém elementu nebo atributu se vyskytují.

Filtrování stránek. Uživateli je nabídnut seznam, jeho zadaných webových adres se všemi podstránkami nacházející se na daných adresách. Tyto podstránky jsou porovnávány na „základní“ podobnost dvou stránek. Zde si může uživatel vybrat, které podstránky projdou dále k hlubšímu porovnávání a které spolu porovnávány nebudou.

Podstránky, které prošly filtrem, jsou porovnávány na, podle uživatelem navoleného výpočtu, podobnost dvou stránek. K těmto porovnávaným podstránkám je také předložen seznam klíčových slov s jejich výskytem. Všechny porovnávané podstránky nebo slova jsou také porovnávány pro celý web. Výsledkem tohoto porovnání je podobnost konkrétních dvou webů, podle zadaného výpočtu.

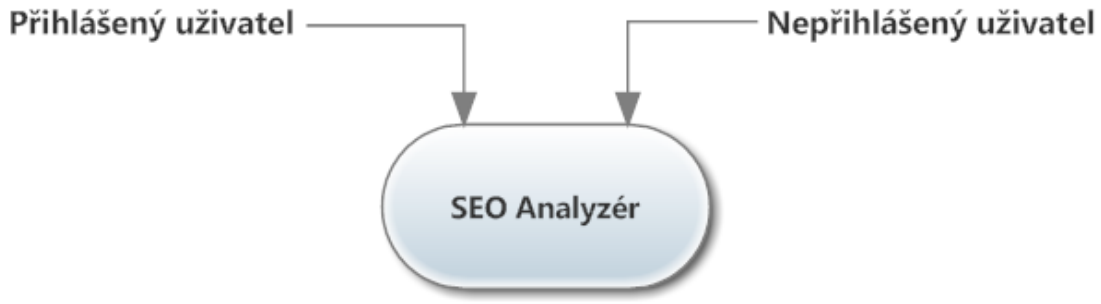
3.2.2 Vstupní data IS

- **Webové adresy** – referenční web, adresa webu, procházení subdomén
- **Respektovaní roboti** – respektované pravidla v robots.txt, respektování atributu rel
- **Vytváření n-gramů** – velikost n-gramů, minimální velikost slova, povolení čísel, seřazení n-gramů podle abecedy, rozlišování diakritiky
- **Váhy jednotlivých HTML elementů a atributů** – významnost elementů a atributů
- **Filtr podobnosti stránek** – minimální výskyt klíčového slova v dokumentu před vážením, minimální výskyt po vážení, minimální podobnost dvou stránek vyjádřená v procentech nebo v četnosti
- **Způsob vypočítání výsledku** – podíl u jednotlivých klíčových slov, podíl u celé stránky a domény, výpočet podobnosti dvou domén
- **Potvrzení a modifikování předloženého výsledku z filtru stránek** – tento výsledek může uživatel zmanipulovat, pomocí vybrání nebo odebrání dvojice stránek které chce nebo nechce porovnávat

3.2.3 Výstupní data IS

- **Výsledek filtru stránek** – výsledkem je podobnost všech stránek na dvou konkurenčních doménách
- **Podobnost dvou domén** – zprůměrovaná podobnost všech stránek, které prošly filtrem, na dvou konkurenčních doménách, nebo podobnost všech slov, stránek prošlých filtrem, na těchto dvou konkurenčních doménách
- **Podobnost stránek na konkurenčních doménách** – vzájemná podobnost stránek konkurenčních domén, dle zadaných kritérií pro výpočet podílu
- **Podobnost klíčových slov na konkurenčních stránkách** – výpis, výskyt a podobnost všech klíčových slov u dvou konkurenčních stránek na dvou doménách

3.2.4 Okolí IS



Obrázek 4: Kontextový diagram

3.2.5 Role uživatelů

Přihlášený uživatel

- Nastavení výše zmíněných kritérií, pro procházení stránek a výpočet výsledku
- Modifikace výsledku, předloženého filtrem
- Uložení nadefinovaného nastavení pro pozdější použití, ať už stejným uživatelem, nebo jiným
- Uložení výsledku porovnávání webů, stránek a klíčových slov, pro pozdější porovnání

Nepřihlášený uživatel

- Nastavení výše zmíněných kritérií, pro procházení stránek a výpočet výsledku
- Modifikace výsledku, předloženého filtrem

3.3 Datová analýza IS

Obsahuje lineární zápis typu entit, E-R konceptuální model a datový slovník. Slouží především k analyzování a popisu struktury a komunikace databáze.

3.3.1 Lineární zápis typu entit

Představuje základní specifikaci jednotlivých typů entit a vztahů mezi nimi, které jsou realizovány prostřednictvím klíčových atributů a cizích klíčů. Podrobnější specifikaci jednotlivých typů entit, je možno nalézt v datovém slovníku a vztahy mezi entitami v E-R konceptuálním modelu.

Ukázka lineárního zápisu u dvou tabulek. První obsahuje základní rozpoznávací faktory jednotlivých stránek a webů v síti a druhá tabulka obsahuje obsah rozparsovaného elementu „a“ a jeho atributů.

analyzer_urlAndConten	**web_id**, **page_id**, url, language, response_code, content, content_md5, all_subdomains, datetime
analyzer_parser_a	**web_id**, **page_id**, **a_id**, href, relative, anchor, image, title

Legenda: **primární klíč**, **cizí klíč**

3.3.2 E-R konceptuální model

Hlavním úkolem konceptuálních modelů, je nalezené entit, vztahů a atributů. E-R model nám pomáhá, na konceptuální úrovni abstrakce, popsat uživatelskou aplikaci za účelem specifikovat strukturu databáze. E-R model tedy z velké části ovlivňuje kvalitu a rychlost aplikace, ať už desktopové nebo webové.

(E-R konceptuální model, je uveden v příloze.)

3.3.3 Datový slovní

Datový slovník, slouží k popisu atributů jednotlivých typů entit. Popisuje jejich funkci, datový typ, povinnost, výchozí hodnotu atd.

Ukázka datového slovníku obsahující výše propagované tabulky v lineárním slovníku (Tabulka 3).

analyzer_dictionarySorted					
Atribut	**Datový typ**	**Primární klíč**	**Cizí klíč**	**Povinný**	**Výchozí hodnota**
ngram_id	BIGINT	ANO	NE	ANO	
ngram_size	TINYINT	ANO	NE	ANO	
ngram	VARCHAR(150)	NE	NE	ANO	

Tabulka 3: Ukázka – Datový slovník

analyzer_urlAndContent					
Atribut	**Datový typ**	**Primární klíč**	**Cizí klíč**	**Povinný**	**Výchozí hodnota**
web_id	BIGINT	ANO	NE	ANO	
page_id	SMALLINT	ANO	NE	ANO	
url	TINYTEXT	NE	NE	ANO	
language	VARCHAR(20)	NE	NE	NE	
response_code	SMALLINT	NE	NE	ANO	
content	MEDIUMTEXT	NE	NE	NE	
content_md5	VARCHAR(32)	NE	NE	NE	
all_subdomains	BOOLEAN	NE	NE	ANO	
datetime	DATETIME	NE	NE	ANO	

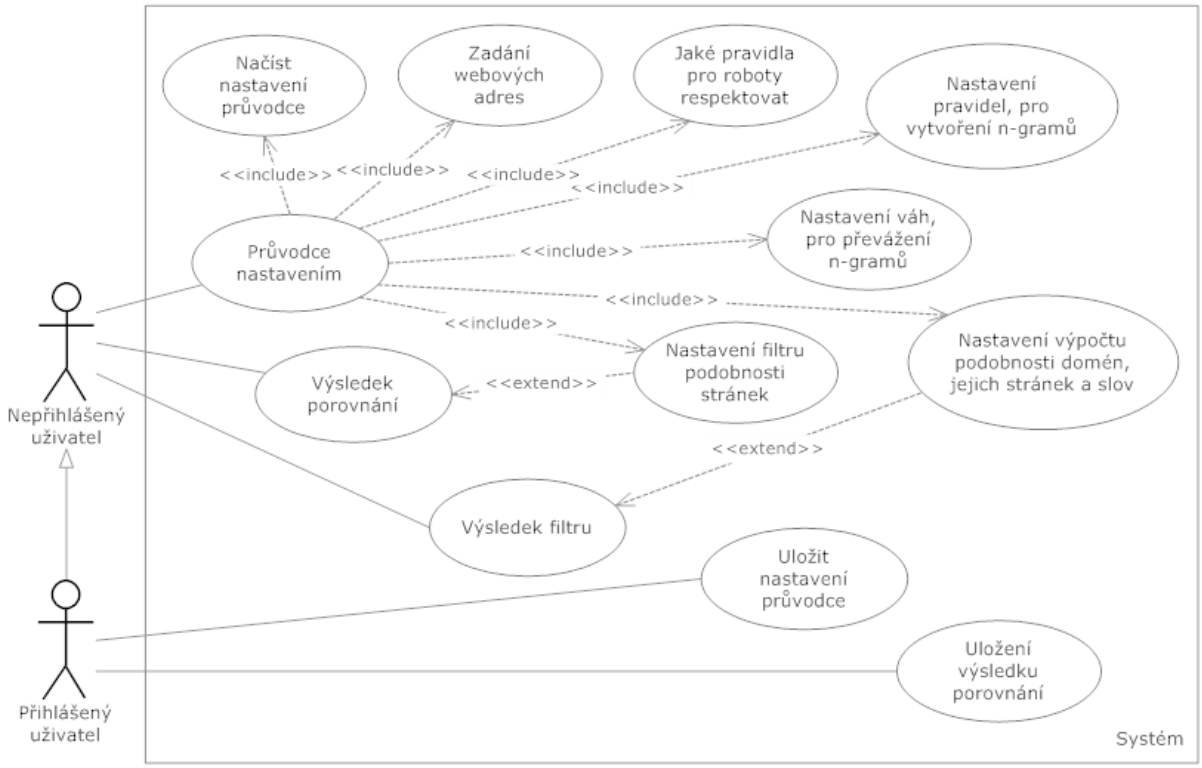
Tabulka 4: Ukázka – Datový slovník

3.4 Analýza procesů IS

Je klíčovým prvkem při tvorbě IS. Při této analýze je nutné identifikovat klíčové procesy, které jsou nezbytné pro správnou funkčnost daného IS. Tato analýze je obvykle založena na konzultaci mezi zákazníkem a konzultantem.

3.4.1 Use-case diagram

Use-case diagram je zobrazení struktury IS z pohledu uživatele. Primárně je určen k definici chování IS, bez toho aniž by odhaloval vnitřní strukturu IS (Obrázek 5).



Obrázek 5: Use-case diagram

3.4.2 Sekvenční diagram

Sekvenční diagram popisuje, jak objekty navzájem mezi sebou komunikují určité časové posloupnosti. Patří do skupiny *diagramů interakce*. Tento typ diagramů, popisuje chování objektů v rámci jednoho scénáře.

(Sekvenční diagram je uveden v příloze.)

3.4.3 Diagram aktivit

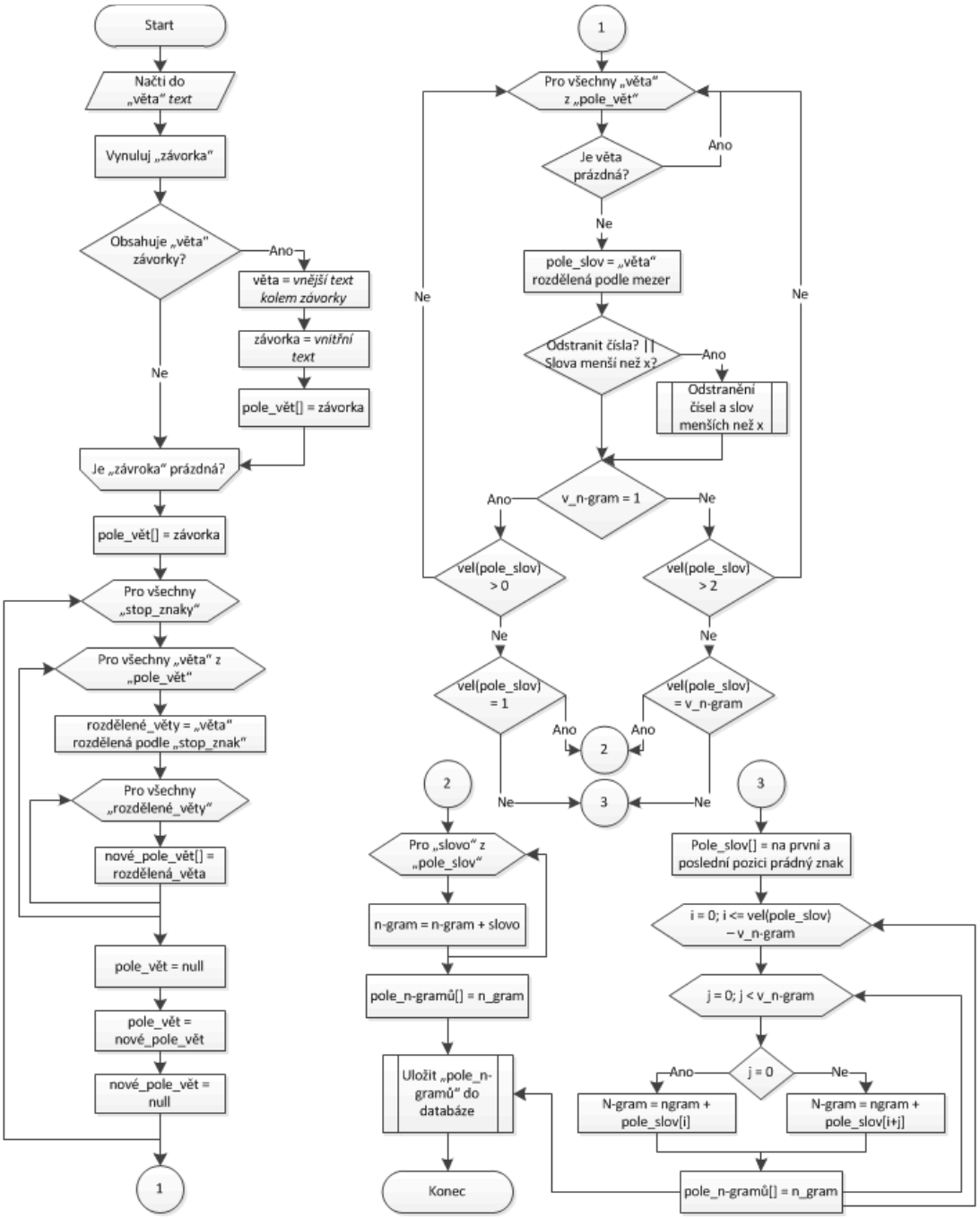
Diagram aktivit je jedním z UML diagramů. Zobrazuje posloupnost aktivit, které mohou probíhat jak sekvenčně, tak paralelně. Využívá se při popisu procedurální logiky, modelování logiky scénářů případů užití, či modelování pracovních byznys procesů.

(Diagram aktivit je uveden v příloze.)

3.4.4 Vývojový diagram

Vývojové diagramy znázorňují průběh či stavbu programu, je to tedy grafické znázornění algoritmu. Algoritmus je přesný postup, který vede k vyřešení určitého problému.

Uvedený vývojový diagram (Obrázek 6) znázorňuje grafickou podobu algoritmu IS, který vytváří a ukládá n-gramy do databáze.



Obrázek 6: Vývojový diagram

3.5 Popis implementace webové aplikace

Následující část této práce popisuje způsob implementace. Jsou zde obsaženy informace o použitých technologiích, vývojových nástrojích a samotné implementaci webové aplikace.

3.5.1 Návrh implementace

V předchozím kroku byl vysvětlen a zprostředkován návrh k tomuto IS. Nyní je řada na samotné implementaci IS. Prvním krokem, je volba programovacího jazyku a SŘBD.

Pro tvorbu webových aplikací, existuje mnoho kvalitních programovacích jazyků. Mezi ty nejznámější patří především PHP [15], Java a .NET. Všechny tyto technologie jsou schopny pracovat na straně serveru, což je pro náš IS stěžejní. Nevýhodou technologií Java a .NET je malá serverová podpora a mnohdy taky nekvalitní. Na druhou stranu, nevyvratitelnou výhodou oproti PHP je výpočetní výkon těchto dvou jazyků, který je výrazně vyšší. PHP ve výpočetním výkonu před jazyky Java a .NET zaostává především pro to, že je to skriptovací programovací jazyk. Nicméně PHP má také výhodu, kterou je jednoduchost tohoto jazyku a mnohem větší serverová podpora. Z důvodu větší serverové podpory, byl pro tuto práci zvolen právě programovací jazyk PHP, jelikož systém bude nasazen v praxi na serveru, který podporuje PHP 5. Na tomto serveru, bude IS spolupracovat s ostatními IS k dosažení nejlepších výsledku co se kvantity i kvality týče.

Ve výběru SŘBD bylo nutné zohlednit, aby bylo možné systém používat i v komerčním projektu. V této situaci se nabízí jako nejlepší řešení zvolit MySQL [16], které je možno používat i v komerčním projektu. MySQL je spolehlivým SŘBD, který je pravidelně aktualizován a spravován. Avšak oproti například SŘBD Oracle zaostává ve výpočetním výkonu.

Pro webovou aplikaci byl tedy zvolen skriptovací programovací jazyk PHP 5 a SŘBD MySQL.

3.5.2 Implementace a ladění

Pro testování a přípravu programu byl zvolen LAMP server [19], tedy Linux, Apache, MySQL a PHP. Je tedy zřejmé, že IS běží na operačním systému Linux, který patří mezi Unixové systémy. Konkrétně je zvoleným systémem Ubuntu, které běží na Linuxovém jádře. Apache je už samotný server, který existuje i ve verzi pro Windows, ale většina internetových serverů běží na Linuxovém serveru. Co se týče MySQL a PHP, to je již objasněno výše.

Samotná implementace programu proběhla ve vývojovém prostředí NetBeans [18], který je zdarma a podporuje nejen PHP, ale také Javu.

Pro návrh kontextového, use-case a sekvenčního diagramu byl použit program SmartDraw[20], který je světově nejpoužívanějším programem svého druhu. Jeho nevýhodou je nutnost zakoupení po uplynutí časové doby určené k vyzkoušení programu. K vytvoření vývojové diagramu byl použit jiný nástroj a to Microsoft Visio 2010[21]. U tohoto nástroje je opět nutností jeho zakoupení, ale Microsoft nabízí studentům VŠB licenci zdarma, čehož jsem využil. Posledním nástrojem k návrhu E-R

konceptuálního modelu databáze byl použit Workbench [17], který je určen přímo pro návrh MySQL databázi. Tento nástroj je produktem stejné firmy, která spravuje SŘBD MySQL, je tedy zaručena přesnoust návrhu se skutečnými hodnotami a možnostmi databáze.

3.5.3 Ukázka SQL dotazu pro porovnání dvou stránek

Následující ukázkou z IS je SQL dotaz, který vybírá všechny slova dvou stránek, tedy referenční a konkurenční, a porovnává jejich podobnost. Dotaz filtruje slova, podle toho jaký minimální výskyt si uživatel nastavil, jakou velikost n-gramu a vybírá slova ze slovníku, podle nastavení, tedy seřazený/neseřazený, s češtinou/bez češtiny.

```
( SELECT rr.ngram_id, rr.count, ro.count, ro.ngram_id, d.ngram
FROM
      ( ( `„Konkrétní tabulka s nepřeváženými hodnotami n-gramů“` AS r JOIN
`„Konkrétní tabulka s převáženými hodnotami n-gramů“` AS rr ON ( r.ngram_id
= rr.ngram_id
                  AND r.web_id = „ID referenčního webu“
                  AND rr.web_id = r.web_id
                  AND r.page_id = „ID stránky referenčního webu“
                  AND rr.page_id = r.page_id
                  AND r.ngram_size = „Velikost n-gramů“
                  AND rr.ngram_size = r.ngram_size
                  AND r.description + r.keywords + r.title + r.h1 + r.h2 +
            r.h3 + r.h4 + r.h5 + r.h6 + r.a_href_domain + r.a_href_url +
            r.a_href_anchor + r.a_href_title + r.p + r.strong + r.em + r.b
            + r.i + r.li + r.dd + r.dt + r.img_src_domain + r.img_src_url +
            r.img_src_alt + r.img_src_title + r.domain + r.url >=
            „Minimální váha před vážením“
                  AND r.rule_id = „ID pravidel pro vytvoření n-gramů“
                  AND rr.weight_id = „ID váh pro převážení n-gramů“
                  AND rr.count >= „Minimální váha po vážení“ ) )
            JOIN `„Konkrétní slovník s frázemi v databázi“` AS d
                  ON r.ngram_id = d.ngram_id )
      LEFT JOIN
            ( `„Konkrétní tabulka s nepřeváženými hodnotami n-gramů“` AS o
      JOIN „Konkrétní tabulka s převáženými hodnotami n-gramů“` AS ro ON (
      o.ngram_id = ro.ngram_id
                  AND o.web_id = „ID konkurenčního webu“
                  AND ro.web_id = o.web_id
                  AND o.page_id = „ID stránky konkurenčního webu“
                  AND ro.page_id = o.page_id
                  AND o.ngram_size = „Velikost n-gramů“
                  AND ro.ngram_size = o.ngram_size
                  AND r.description + r.keywords + r.title + r.h1 + r.h2 +
            r.h3 + r.h4 + r.h5 + r.h6 + r.a_href_domain + r.a_href_url +
            r.a_href_anchor + r.a_href_title + r.p + r.strong + r.em + r.b
            + r.i + r.li + r.dd + r.dt + r.img_src_domain + r.img_src_url +
```

```
            r.img_src_alt + r.img_src_title + r.domain + r.url >=
            „Minimální váha před vážením“
                  AND o.rule_id = „ID pravidel pro vytvoření n-gramů“
                  AND ro.weight_id = „ID váh pro převážení n-gramů“
                  AND ro.count >= „Minimální váha po vážení“ ) )
      ON r.ngram_id = o.ngram_id )

UNION

( SELECT rr.ngram_id, rr.count, ro.count, ro.ngram_id, d.ngram
FROM
      ( `„Konkrétní tabulka s nepřeváženými hodnotami n-gramů“` AS r JOIN
`„Konkrétní tabulka s převáženými hodnotami n-gramů“` AS rr ON ( r.ngram_id
= rr.ngram_id
                  AND r.web_id = „ID referenčního webu“
                  AND rr.web_id = r.web_id
                  AND r.page_id = „ID stránky referenčního webu“
                  AND rr.page_id = r.page_id
                  AND r.ngram_size = „Velikost n-gramů“
                  AND rr.ngram_size = r.ngram_size
                  AND r.description + r.keywords + r.title + r.h1 + r.h2 +
            r.h3 + r.h4 + r.h5 + r.h6 + r.a_href_domain + r.a_href_url +
            r.a_href_anchor + r.a_href_title + r.p + r.strong + r.em + r.b
            + r.i + r.li + r.dd + r.dt + r.img_src_domain + r.img_src_url +
            r.img_src_alt + r.img_src_title + r.domain + r.url >=
            „Minimální váha před vážením“
                  AND r.rule_id = „ID pravidel pro vytvoření n-gramů“
                  AND rr.weight_id = „ID váh pro převážení n-gramů“
                  AND rr.count >= „Minimální váha po vážení“) )
      RIGHT JOIN
                  ( ( `„Konkrétní tabulka s nepřeváženými hodnotami n-
            gramů“` AS o JOIN `„Konkrétní tabulka s převáženými hodnotami
            n-gramů“` AS ro ON ( o.ngram_id = ro.ngram_id
                  AND o.web_id = „ . $this->webID . “
                  AND ro.web_id = o.web_id
                  AND o.page_id = „ . $pageID . “
                  AND ro.page_id = o.page_id
                  AND o.ngram_size = „Velikost n-gramů“
                  AND ro.ngram_size = o.ngram_size
                  AND r.description + r.keywords + r.title + r.h1 + r.h2 +
            r.h3 + r.h4 + r.h5 + r.h6 + r.a_href_domain + r.a_href_url +
            r.a_href_anchor + r.a_href_title + r.p + r.strong + r.em + r.b
            + r.i + r.li + r.dd + r.dt + r.img_src_domain + r.img_src_url +
            r.img_src_alt + r.img_src_title + r.domain + r.url >=
            „Minimální váha před vážením“
                  AND o.rule_id = „ID pravidel pro vytvoření n-gramů“
                  AND ro.weight_id = „ID váh pro převážení n-gramů“
```

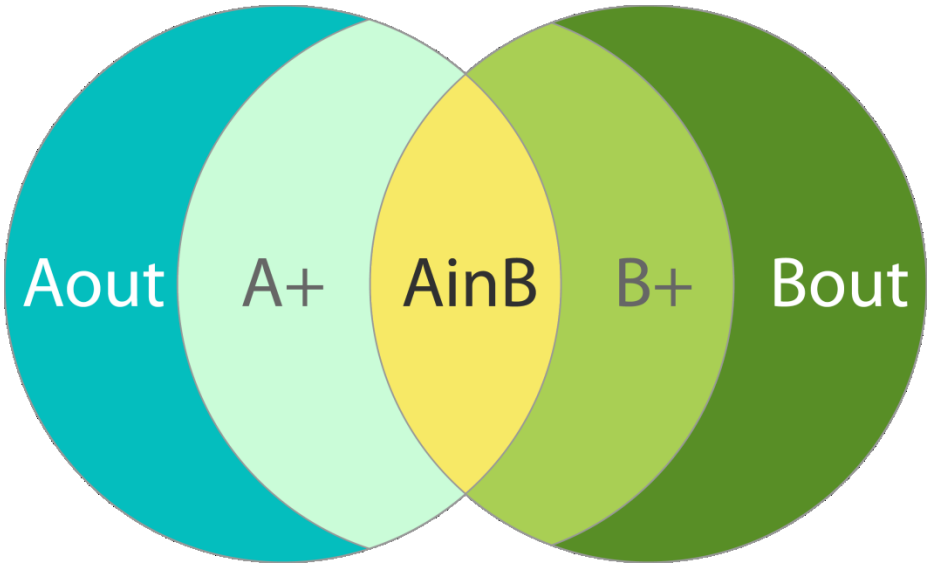
```
            AND ro.count >= „Minimální váha po vážení“) )
      JOIN `„Konkrétní slovník s frázemi v databázi“` AS d
            ON o.ngram_id = d.ngram_id )
ON r.ngram_id = o.ngram_id );
```

3.6 Vyhodnocení nad shromážděnými daty

IS je vytvořen tak, aby si uživatel mohl zvolit z různých kombinací pro porovnávání shromážděných dat. U porovnávání jednotlivých klíčových slov, kdy je každé klíčové slovo jedné stránky porovnáváno se svým protějškem na druhé stránce, má uživatel na výběr ze tří možností porovnávání. U porovnávání dvou domén a taky u porovnávání dvou stránek má uživatel na výběr z celkem 1013 různých kombinací.

3.6.1 Porovnání dvou různých domén a dvou stránek

Uživatel může různě kombinovat podíly klíčových slov, které mají domény/stránky společné, společné ale jedna stránka jich má více/méně a slova, které se na jedné stránce vyskytují a na druhé ne. Následující obrázek (Obrázek 7) představuje množinu všech klíčových slov, které se vyskytují na dvou porovnávaných doménách/stránkách.



Obrázek 7: Množina klíčových slov, na dvou doménách/stránkách

Aout	klíčové slova webu A, které se nenachází na webu B
A+	klíčové slova webu A, nacházející se na webu A ve větším množství, než na webu B
AinB	klíčové slova vyskytující se na webu A i B ve stejném, nenulovém množství
B+	klíčové slova webu B, nacházející se na webu B ve větším množství, než na webu A
Bout	klíčové slova webu B, které se nenachází na webu A
ABall	množina všech klíčových slov, vyskytujících se na doménách/stránkách A a B, tedy: „Aout“ + „A+“ + „AinB“ + „B+“ + „Bout“

Tabulka 5: Legenda k množině klíčových slov, na dvou doménách/stránkách

Všechny tyto slova, může uživatel přiřadit jak do čitatele, tak do jmenovatele zlomku, který vypočítává podobnost dvou domén/stránek. Zlomek vypočítávající například kolik klíčových slov, mají stránky společných z celkového počtu slov, bude vypadat následovně:

$$\frac{(A+)+(AinB)+(B+)}{(Aout)+(A+)+(AinB)+(B+)+(Bout)}$$

Je zřejmé, že si uživatel takto může navolit z mnoha možností výpočtu, ať už je jejich výsledek relevantní, nebo ne. Přesný počet kombinací počítá následující rovnice:

$$\binom{10}{2}+\binom{10}{3}+\binom{10}{4}+\binom{10}{5}+\binom{10}{6}+\binom{10}{7}+\binom{10}{8}+\binom{10}{9}+\binom{10}{10}=1013$$

3.6.2 Porovnání výskytu klíčového slova na dvou stránkách

Při porovnávání klíčových slov je situace jednodušší. Je zde na výběr ze tří možností pro výpočet podobnosti, a to:

- $\frac{Referenční\ web}{Konkurenční\ web}$
- $\frac{Konkurenční\ web}{Referenční\ web}$
- $\frac{Větší\ výskyt\ slova}{Menší\ výskyt\ slova}$

Nejsnadnější je vysvětlení na příkladu. Máme tedy příklad, kde:

- Jako referenční web je zvolen web A a jako konkurenční web B
- Reálné klíčové slova, reprezentují znaky m, n, o, p, q

Výskyt na webu A	Klíčové slovo	Výskyt na webu B	Závěr
0	**m**	2	Bout
1	**n**	4	B+
2	**o**	2	AinB
5	**p**	3	A+
3	**q**	0	Aout

Tabulka 6: Příklad - Výskyt frází na stránkách A a B

Zvolený přepočet:

Klíčové slovo	Referenční/Konkurenční	Konkurenční/Referenční	Větší/Menší
m	-	$\frac{2}{0}$	$\frac{2}{0}$
n	$\frac{1}{4}$	$\frac{4}{1}$	$\frac{4}{1}$
o	$\frac{2}{2}$	$\frac{2}{2}$	$\frac{2}{2}$
p	$\frac{5}{3}$	$\frac{3}{5}$	$\frac{5}{3}$
q	$\frac{3}{0}$	-	$\frac{3}{0}$

Tabulka 7: Příklad - Zvolený přepočet u výskytu frází na A a B

Finální tabulka výskytu klíčového slova na stránce vůči jiné stránce:

Klíčové slovo	Referenční/Konkurenční	Konkurenční/Referenční	Větší/Menší
m	-	+	+
n	25%	400%	400%
o	100%	100%	100%
p	166,66%	60%	166,66%
q	+	-	+

Tabulka 8: Příklad: Výsledek porovnání frází na stránkách A a B

3.7 Popis reálného příkladu

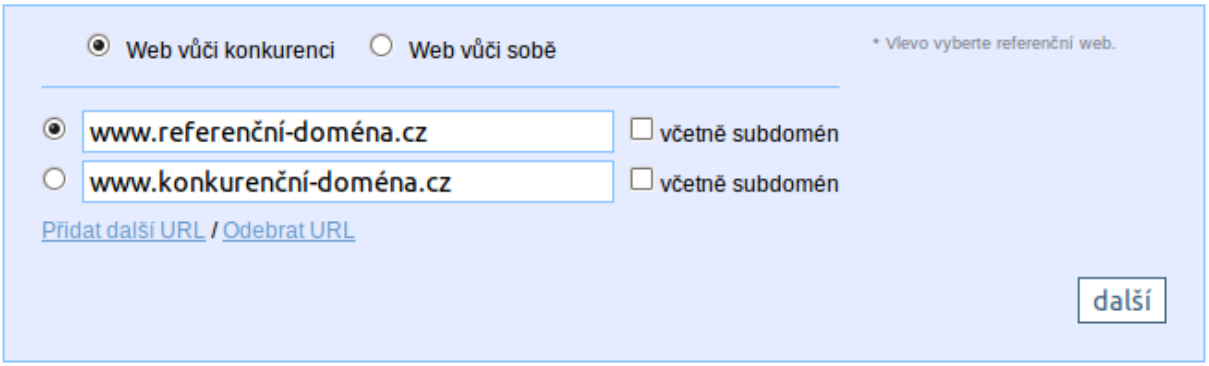
V poslední části, co se tohoto informačního systému týče, je popis, příkladu, jak se systémem pracuje. IS je v podstatě uživatelsky rozdělen do třech hlavních částí, kterými jsou: průvodce nastavením, potvrzení profiltrovaných stránek a výsledek porovnávaných domén, stránek a klíčových slov.

3.7.1 Průvodce nastavením

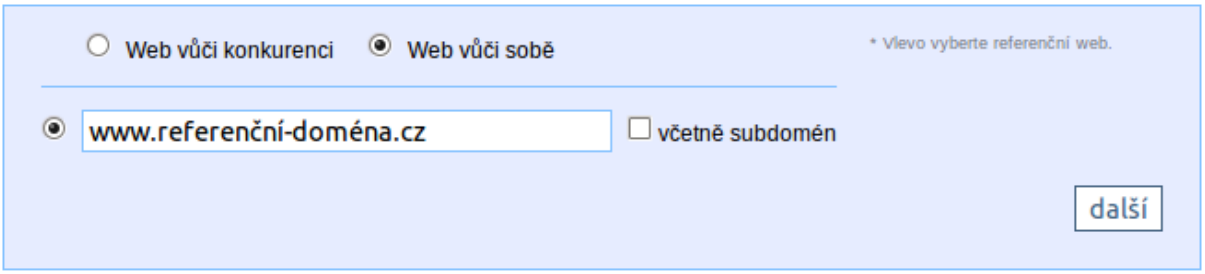
Průvodce nastavením, je rozdělen do šesti kroků, kterými musí uživatel projít.

Zadání webových adres

Je zde na výběr ze dvou možností, které ovlivňují jak zadání stránek, tak průběh porovnávání stránek (Obrázek 8 a 9). Těmi možnostmi jsou, zda skenovat *web vůči konkurenci* nebo *web vůči sobě*. Při skenování webu vůči konkurenci, máme možnost zadat n domén, které budou vzájemně porovnávány. Zároveň musí uživatel vybrat *referenční web*, který bude porovnáván se všemi ostatními doménami. U skenování webu vůči sobě, odpadá možnost zadání většího počtu adres, protože je porovnávána pouze jedna zadaná adresa sama se sebou. U obou možností, si může uživatel zvolit, zda procházet doménu, včetně jeho subdomén, nebo jen zadanou doménu.



Obrázek 8: Formulář - Zadání webových adres (web vůči konkurenci)

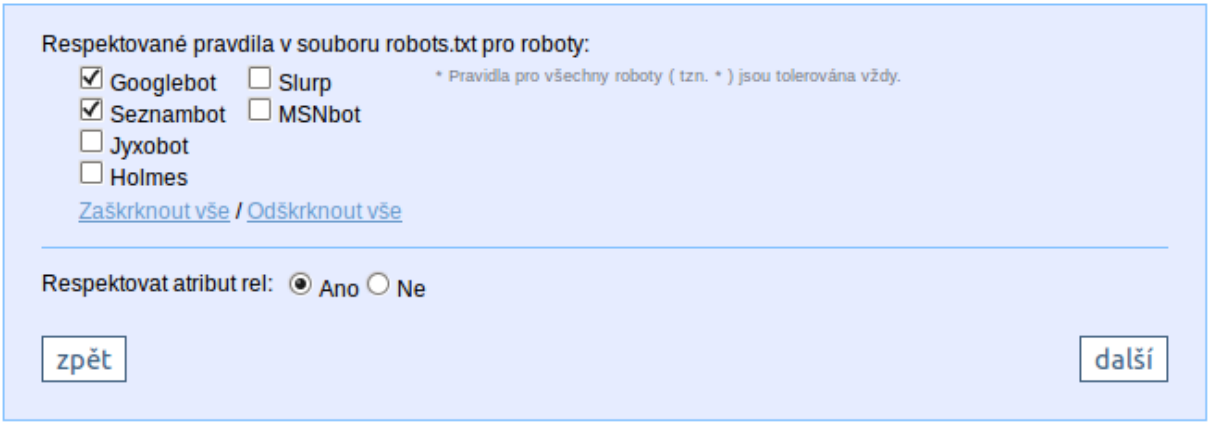


Obrázek 9: Formulář - Zadání webových adres (web vůči sobě)

Vybrání pravidel pro roboty, která budou respektována

V tomto kroku uživatel vybírá, pro které User-agenty ze souboru robots.txt, bude respektovat pravidla. Má na výběr z na českém trhu nejpodstatnějších robotů (Obrázek 10), s tím, že pravidla pro všechny roboty, tedy User-agent označený symbolem *, budou respektována vždy, a uživatel to nemůže ovlivnit.

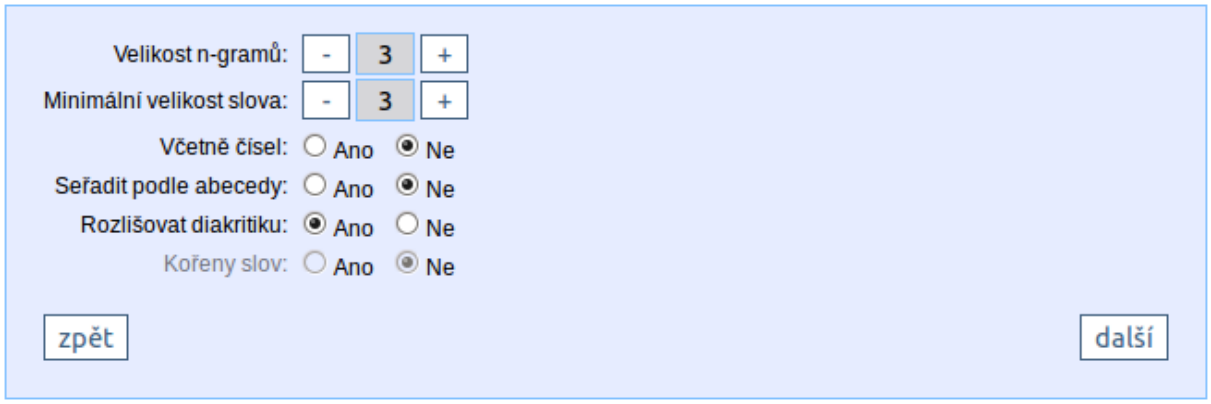
Další možností je, zda bude aplikace respektovat atribut `rel="nofollow"` u elementu `a` (Obrázek 10).



Obrázek 10: Formulář - Výběr respektovaných robotů a atribut rel

Pravidla pro vytvoření n-gramů

Zde si uživatel nastaví *velikost n-gramů*, *minimální velikost slova*, zda propouštět filtrem *čísla*, *seřazení* n-gramů podle abecedy a zda bude aplikace rozlišovat slova s *diakritikou* a bez diakritiky (Obrázek 11).



Obrázek 11: Formulář - Pravidla pro vytvoření n-gramů

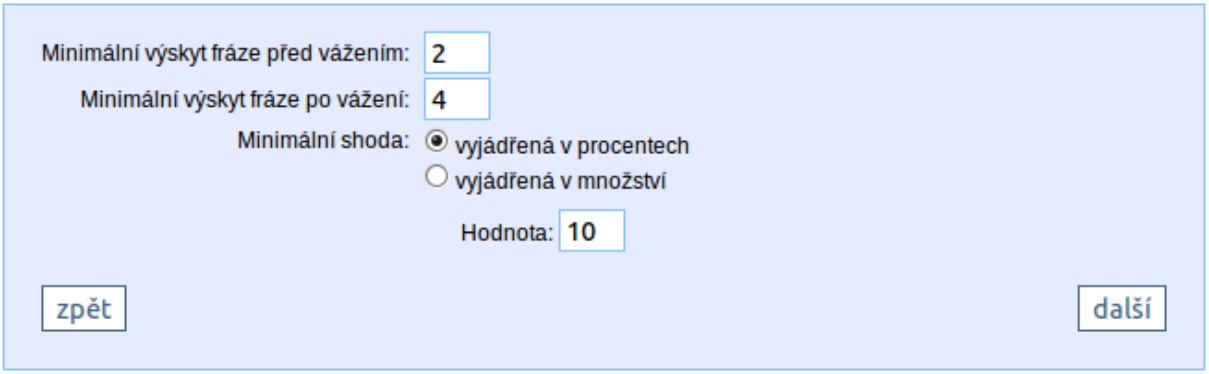
Převážení n-gramů

Dalším krokem je nastavení důležitosti ke každému elementu nebo atributu elementu. To je uskutečněno nastavením vah jednotlivým HTML tagům (Obrázek 12). V aplikaci je poté každý element pře násobenou touto hodnou, takže se na stránce nevyskytuje například jeden krát, ale n krát.

Obrázek 12: Formulář - Nastavení vah u HTML tagů

Filtr podobnosti stránek

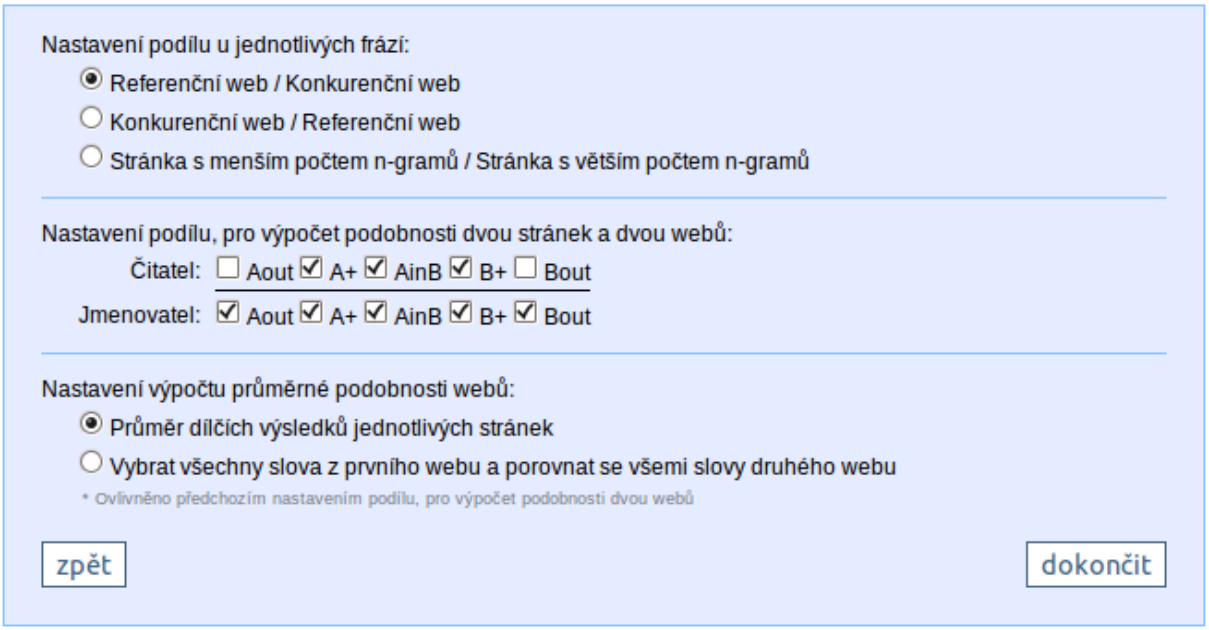
Dále uživatel nastaví kritéria filtru, která ovlivňují rozhodnutí aplikace, zda se dané dvě stránky podobají, nebo ne. Tento krok je také ovlivněn krokem 1, kde se při porovnávání webu vůči sobě, dvě stránky s podobností 100% nepropouští přes filtr. Kritéria, která uživatel nastavuje, jsou: *minimální výskyt klíčového slova před vážením* HTML tagů, *minimální výskyt klíčového slova po vážení* HTML tagů, jak zobrazovat shodu dvou stránek (procento, či četnost) a jaká je tato minimální shoda (Obrázek 13).



Obrázek 13: Formulář - Filtr podobnosti stránek

Způsob vypočítání výsledku podobnosti

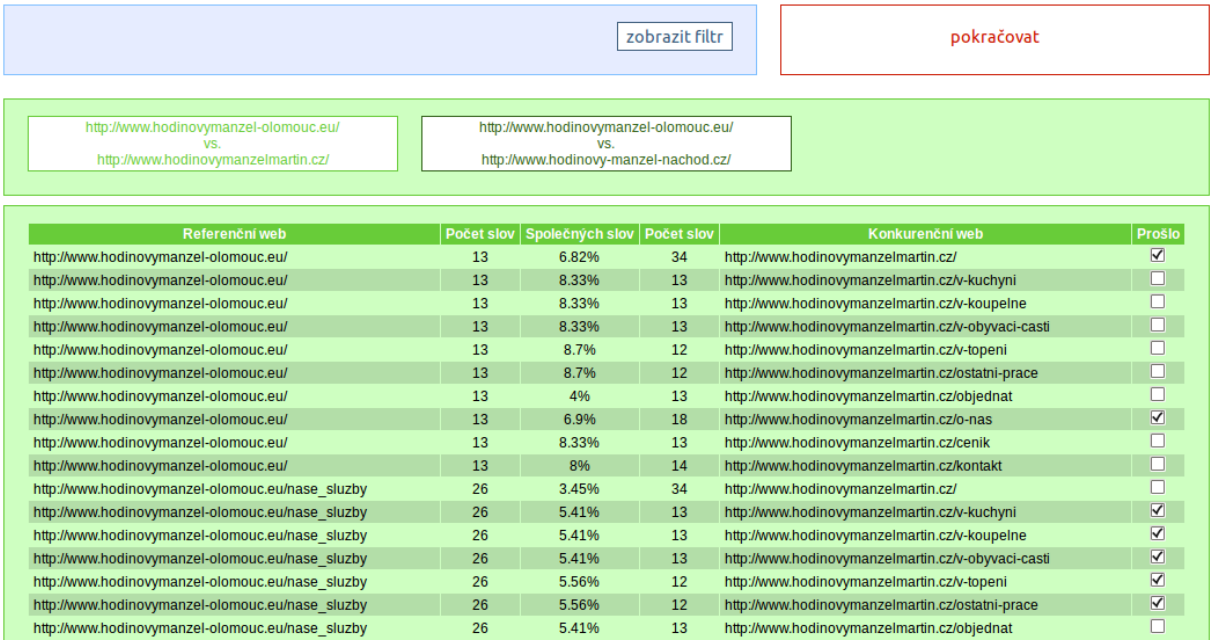
Posledním krokem v průvodci nastavením je nastavení kritérií, pro výpočet podobnosti dvou domén, stránek a klíčových slov (Obrázek 14). U všech těchto možností se nastavuje, jak bude vypadat podíl dvou porovnávaných hodnot. U porovnávání *klíčových slov*, má uživatel na výběr mezi třemi druhy zlomku a u porovnávání *domén a stránek* má uživatel na výběr z 1013 různých kombinací. Poslední možností je, zda se bude podobnost dvou domén vypočítávat z *dílčích výsledků* podobnosti stránek nebo budou vybrány *všechny slova* na těchto porovnávaných doménách a vloží se do zvoleného zlomku.



Obrázek 14: Formulář - Způsob vypočítání podobnosti

3.7.2 Potvrzení nebo modifikace filtru

Po projití, rozparsování, vytvoření a převážení n-gramů je uživateli nabídnut výsledek filtru, pro zkontrolování, zda výsledek souhlasí s reálem, nebo s tím, co chce uživatel porovnávat. V tomto kroku má uživatel možnost, *změnit nastavení filtru*, který nastavoval v pátém kroku průvodce nastavením. Další možnosti je, zvolit pomocí chceckboxu na pravé straně tabulky s vypsanými výsledky, které stránky mají být označené jako prošlé filtrem a které nikterak.



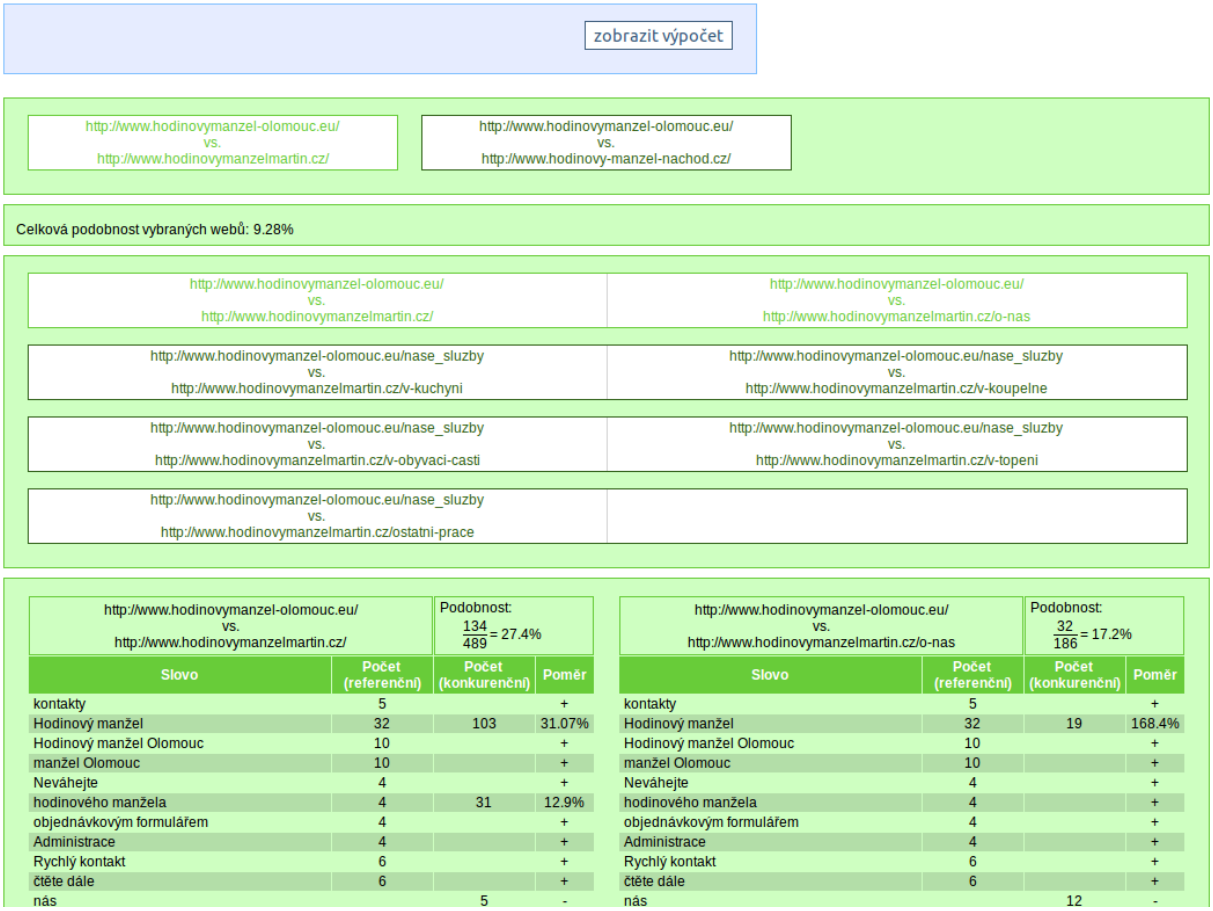
Referenční web	Počet slov	Společných slov	Počet slov	Konkurenční web	Prošlo
http://www.hodinovymanzel-olomouc.eu/	13	6.82%	34	http://www.hodinovymanzelmartin.cz/	☑
http://www.hodinovymanzel-olomouc.eu/	13	8.33%	13	http://www.hodinovymanzelmartin.cz/v-kuchyni	☐
http://www.hodinovymanzel-olomouc.eu/	13	8.33%	13	http://www.hodinovymanzelmartin.cz/v-koupelne	☐
http://www.hodinovymanzel-olomouc.eu/	13	8.33%	13	http://www.hodinovymanzelmartin.cz/v-obyvaci-casti	☐
http://www.hodinovymanzel-olomouc.eu/	13	8.7%	12	http://www.hodinovymanzelmartin.cz/v-topeni	☐
http://www.hodinovymanzel-olomouc.eu/	13	8.7%	12	http://www.hodinovymanzelmartin.cz/ostatni-prace	☐
http://www.hodinovymanzel-olomouc.eu/	13	4%	13	http://www.hodinovymanzelmartin.cz/objednat	☐
http://www.hodinovymanzel-olomouc.eu/	13	6.9%	18	http://www.hodinovymanzelmartin.cz/o-nas	☑
http://www.hodinovymanzel-olomouc.eu/	13	8.33%	13	http://www.hodinovymanzelmartin.cz/cenik	☐
http://www.hodinovymanzel-olomouc.eu/	13	8%	14	http://www.hodinovymanzelmartin.cz/kontakt	☐
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	3.45%	34	http://www.hodinovymanzelmartin.cz/	☐
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	5.41%	13	http://www.hodinovymanzelmartin.cz/v-kuchyni	☑
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	5.41%	13	http://www.hodinovymanzelmartin.cz/v-koupelne	☑
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	5.41%	13	http://www.hodinovymanzelmartin.cz/v-obyvaci-casti	☑
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	5.56%	12	http://www.hodinovymanzelmartin.cz/v-topeni	☑
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	5.56%	12	http://www.hodinovymanzelmartin.cz/ostatni-prace	☑
http://www.hodinovymanzel-olomouc.eu/nase_sluzby	26	5.41%	13	http://www.hodinovymanzelmartin.cz/objednat	☐

Obrázek 15: Formulář - Potvrzení a modifikace hodnot filtru podobnosti stránek

3.7.3 Zobrazení výsledku

V posledním kroku, je již samotné zobrazení výsledku. Uživatel má možnost, stejně jako v předchozím kroku, *změnit nastavení výpočtu*, které nastavoval v šestém kroku průvodce nastavením.

Na této stránce uživatel nalezne podobnost domén s referenční doménou, jejich stránek a klíčových slov, které se na těchto stránkách vyskytují.



Obrázek 16: Formulář - Zobrazení výsledku

4. Zhodnocení dosažených výsledku

V porovnání s konkurenčními projekty, které jsou zdarma, jako jsou třeba *http://www.seoworkers.com*, *http://www.seo-analyzator.cz* apod., je v tom, že tyto nástroje pracují pouze tak, že uživateli řeknou, co je na zadané stránce špatně a co je dobře. Což je bezpochyby taky důležitá informace, ale v konkurenčním boji dvou domén, není až tak platná. Oproti tomu, tato aplikace nabízí hodnoty referenčního webu a k nim hodnoty konkurenčního webu. Což se dá mnohem lépe využít, dá se říci, že uživatele tento nástroj navádí k tomu, aby kopíroval konkurenční stránku.

Co do pohledu na stránku týče, můžou být také užitečné nástroje jako *http://www.webseoanalytics.com/free/seo-tools/web-seo-analysis.php*, které ovšem opět nenabízejí pohled na konkurecni, ale dokáží nám zanalizovat stránku, říct, zda klíčové slova na stránce souhlasí s titlke, klik je na stránce klíčových apod.

Dalšími nástroji, které kontrolují konkurenční stránky, jsou nástroje typu *https://adwords.google.com/select/KeywordToolExternal*, které nabízejí pohled na statistiku vyhledávání klíčového slova za určitý časový interval a zároveň, jak moc je velká konkurence u daného klíčového slova. Tento fakt se dá už dobře využít, ale především při výběru, na které slovo se má stránka optimalizovat a ne v tom, jak zlepšit danou stránku na optimalizaci konkrétního, již vybraného a částečně optimalizovaného, slova.

Nakonec se dostávám k nástrojům, které dělají právě to, co nabízí tento IS. Mezi takovéto nástroje patří například *http://www.spyfu.com*. Tyto nástroje nabízí porovnání webu s konkurenčními stránkami, co se podobnosti a výskytu klíčových slov týče, což je silný nástroj při závodění na poli SEO optimalizace. Nevýhodou takových to nástrojů je fakt, že jde většinou a zahraniční aplikace a tudíž nepodporují českou diakritiku. To může do velké míry ovlivnit výsledek testu, který aplikace provádí. Naproti tomu, IS, kterým se zabývá tato práce, je zaměřeny právě na české prostředí.

5. Závěr

Hlavním cílem této práce, bylo vytvořit aplikaci pro SEO analyzování jednoho webu vůči konkurenci, z pohledu klíčových slov. Tento systém je vhodný zejména pro uživatele, kteří se vyskytují vysoko ve výsledcích vyhledávání a potřebují zjistit, proč je někdo stále výše než oni. Tento nástroj jim v tomhle ohledu napomůže optimalizovat stránku z hlediska on-page faktorů, na stejnou, nebo lepší vůči konkurenční stránce, která se ve vyhledávání objevuje výše.

Dalším cílem práce bylo, seznámit se s problematikou SEO optimalizace webových stránek. K tomuto účelu, bylo zapotřebí nastudovat vhodnou literaturu, včetně článků týkajících se této problematiky nacházejících se ve vědeckých publikacích.

Systém je v této fázi schopný stáhnout a převést obsah stránek na klíčová slova. Poté je schopen takto získané klíčové slova porovnávat s ostatními stránkami vyskytující se na internetu. Uživateli poskytne informace o podobnosti domén, jejich stránek a klíčových slov na těchto stránkách se vyskytujících. Do budoucna je plánováno, rozšířit vytváření n-gramů, o porovnávání kořenů slov, čímž by se zmenšil objem dat v databázi a zároveň by byl výsledek relevantnější vzhledem k tomu, jak s textem pracují vyhledávače. Dalším možným vylepšením, které by zpřehlednilo zobrazení výsledku porovnání, je zobrazení porovnání formou Vennových diagramů.

6. Seznam použité literatury

[1] KUBÍČEK, Michal. *Velký průvodce SEO*: *jak dosáhnout nejlepších pozic ve vyhledávačích*. Vyd. 1. Brno: Computer Press, 2008, 318 s. ISBN 978-80-251-2195-5.

[2] KUBÍČEK, Michal a Jan LINHART. *333 tipů a triků pro SEO*: *[sbírka nejlepších technik optimalizace webů pro vyhledávače]*. Vyd. 1. Brno: Computer Press, 2010, 262 s. ISBN 978-80-251-2468-0.

[3] GRAPPONE, Jennifer a Gradiva COUZIN. *SEO*: *search engine optimization : ovládněte SEO a získejte výhodu před konkurencí : optimalizujte své webové stránky pro vyhledávací servery : přiveďte na své stránky zákazníky dříve, než to udělá konkurence*. Vyd. 1. Brno: Computer Press, 2010, 262 s. ISBN 978-80-86815-85-5.

[4] DOMES, Martin a Gradiva COUZIN. *SEO*: *jednoduše*. Vyd. 1. Překlad Roman Skřivánek, Dana Balaštíková. Brno: Computer Press, 2011, 141 s. Naučte se za víkend (Computer Press). ISBN 978-80-251-3456-6.

[5] JANOUCH, Viktor a Gradiva COUZIN. *Internetový marketing*: *prosaďte se na webu a sociálních sítích*. Vyd. 1. Překlad Roman Skřivánek, Dana Balaštíková. Brno: Computer Press, 2010, 304 s. Naučte se za víkend (Computer Press). ISBN 978-80-251-2795-7.

[6] Google PageRank, vysvětlení a odpovědi. *jakpsatweb.cz* [online]. Dostupné z www: <http://www.jakpsatweb.cz/seo/pagerank.html>

[7] Google Indexer Work Google Web Indexing Google Index Search How Google Improve Search Performance How Google Improve Indexing. *GoogleTruths.com* [online]. Dostupné z www: <http://www.googletruths.com/Google/Work/how-google- indexer-works.aspx>

[8] Seznam S-Rank Checker - pagerank.jklir.net. *PAGERANK.jklir.net* [online]. Dostupné z www: <http://pagerank.jklir.net/?p=srank>

[9] O Google: algoritmy, vlastnosti Google, jak optimalizovat. *jakpsatweb.cz* [online]. Dostupné z www: <http://www.jakpsatweb.cz/google.html>

[10] Robots.txt - zakázání přístupu robotům. *jakpsatweb.cz* [online]. Dostupné z www: <http://www.jakpsatweb.cz/robots-txt.html>

[11] NAVRCHOLU.cz: Windows Live Search zamíchal statistikou podílů vyhledávačů - Internet Info. *NAVRCHOLU.cz* [online]. Dostupné z www: <http://www.iinfo.cz/tiskove-centrum/tiskove-zpravy/navrcholu-vyhledavace-kveten/>

[12] Podíl vyhledávačů na českém trhu na konci roku 2010 | Web71. w*eb71.cz* [online]. Dostupné z www: <http://www.web71.cz/clanky/35-podil-vyhledavacu-2010/>

[13] TOPlist – Historie. *TOPlist.cz* [online]. Dostupné z www: <http://www.toplist.cz/stat/?a=history&type=4>

[14] OneStat Website Statistics and website metrics - Press Room. *OneStat.com* [online]. Dostupné z www: <view-source:http://www.onestat.com/html/aboutus_pressbox45-search-phrases.html>

[15] PHP: Hypertext Preprocessor. *php.net* [online]. Dostupné z www: <http://php.net/ >

[16] MySQL :: The world's most popular open source database. *MySQL.com* [online]. Dostupné z www: <http://www.mysql.com/>

[17] MySQL :: MySQL Workbench 5.2. *MySQL.com* [online]. Dostupné z www: <http://www.mysql.com/products/workbench/>

[18] NetBeans. *NetBeans.org* [online]. Dostupné z www: <http://netbeans.org/>

[19] Wiki - Ubuntu Česko. *wiki.ubuntu.cz* [online]. Dostupné z www: <http://wiki.ubuntu.cz/>

[20] SmartDraw - Communicate Visually with the World's First Visual Processor(tm). *SmartDraw.com* [online]. Dostupné z www: <http://www.smartdraw.com/>

[21] Microsoft Visio 2010 - Office.com. *office.microsoft.com* [online]. Dostupné z www: <http://office.microsoft.com/en-us/visio/>

7. Seznam příloh

7.1 Lineární zápis typů entit

analyzer_dictionarySorted	**ngram_id**, **ngram_size**, ngram
analyzer_dictionarySortedCZ	**ngram_id**, **ngram_size**, ngram
analyzer_dictionaryUnsorted	**ngram_id**, **ngram_size**, ngram
analyzer_dictionaryUnsortedCZ	**ngram_id**, **ngram_size**, ngram
analyzer_disallowTemp	**disallow_id**,**web_id**, **page_id**, datetime
analyzer_ngramsRules	**rule_id**, min_length, cant_be_num
analyzer_ngramsSorted	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **rule_id**, description, keywords, title, h1, h2, h3, h4, h5, h6, img_src_alt, a_href_domain, a_href_url, a_href_anchor, a_href_title, p, strong, em, b, i ,li, dd, dt, img_src_domain, img_src_url, img_src_title, domain, url
analyzer_ngramsSortedCZ	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **rule_id**, description, keywords, title, h1, h2, h3, h4, h5, h6, img_src_alt, a_href_domain, a_href_url, a_href_anchor, a_href_title, p, strong, em, b, i ,li, dd, dt, img_src_domain, img_src_url, img_src_title, domain, url
analyzer_ngramsUnsorted	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **rule_id**, description, keywords, title, h1, h2, h3, h4, h5, h6, img_src_alt, a_href_domain, a_href_url, a_href_anchor, a_href_title, p, strong, em, b, i ,li, dd, dt, img_src_domain, img_src_url, img_src_title, domain, url
analyzer_ngramsUnsortedCZ	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **rule_id**, description, keywords, title, h1, h2, h3, h4, h5, h6, img_src_alt, a_href_domain, a_href_url, a_href_anchor, a_href_title, p, strong, em, b, i ,li, dd, dt, img_src_domain, img_src_url, img_src_title, domain, url
analyzer_parser_a	**web_id**, **page_id**, **a_id**, href, relative, anchor, image, title
analyzer_parser_b	**web_id**, **page_id**, **b_id**, b
analyzer_parser_dd	**web_id**, **page_id**, **dd_id**, dd
analyzer_parser_description	**web_id**, **page_id**, description
analyzer_parser_dt	**web_id**, **page_id**, **dt_id**, dt
analyzer_parser_em	**web_id**, **page_id**, **em_id**, em
analyzer_parser_h1	**web_id**, **page_id**, **h1_id**, h1
analyzer_parser_h2	**web_id**, **page_id**, **h2_id**, h2
analyzer_parser_h3	**web_id**, **page_id**, **h3_id**, h3
analyzer_parser_h4	**web_id**, **page_id**, **h4_id**, h4
analyzer_parser_h5	**web_id**, **page_id**, **h5_id**, h5

analyzer_parser_h6	**web_id**, **page_id**, **h6_id**, h6
analyzer_parser_i	**web_id**, **page_id**, **i_id**, i
analyzer_parser_img	**web_id**, **page_id**, **link_id**, src, relative, alt, title
analyzer_parser_keywords	**web_id**, **page_id**, keywords
analyzer_parser_li	**web_id**, **page_id**, **li_id**, li
analyzer_parser_p	**web_id**, **page_id**, **p_id**, p
analyzer_parser_strong	**web_id**, **page_id**, **strong_id**, strong
analyzer_parser_keywords	**web_id**, **page_id**, title
analyzer_recalculatedNgramsSorted	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **weight_id**, count
analyzer_recalculatedNgramsSortedCZ	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **weight_id**, count
analyzer_recalculatedNgramsUnsorted	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **weight_id**, count
analyzer_recalculatedNgramsUnsortedCZ	**web_id**, **page_id**, **ngram_id**, **ngram_size**, **weight_id**, count
analyzer_ngramsUnsorted	**weight_id**, description, keywords, title, h1, h2, h3, h4, h5, h6, a_href_domain, a_href_url, a_href_anchor, a_href_title, p, strong, em, b, i ,li, dd, dt, img_src_domain, img_src_url, img_src_alt, img_src_title, domain, url
analyzer_robots_meta	**web_id**, **page_id**, index, follow
analyzer_robots_txt	**web_id**, **page_id**, disallow, allow, for
analyzer_sitemap	**web_id**, **page_id**, **link**
analyzer_urlAndContent	**web_id**, **page_id**, url, response_code, content, content_md5, all_subdomains, datetime

7.2 Datový slovník

analyzer_dictionarySorted					
Atribut	**Datový typ**	**Primární klíč**	**Cizí klíč**	**Povinný**	**Výchozí hodnota**
ngram_id	BIGINT	ANO	NE	ANO	
ngram_size	TINYINT	ANO	NE	ANO	
ngram	VARCHAR(150)	NE	NE	ANO	

analyzer_dictionarySortedCZ					
Atribut	**Datový typ**	**Primární klíč**	**Cizí klíč**	**Povinný**	**Výchozí hodnota**
ngram_id	BIGINT	ANO	NE	ANO	
ngram_size	TINYINT	ANO	NE	ANO	
ngram	VARCHAR(150)	NE	NE	ANO	

analyzer_dictionaryUnsorted					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
ngram_id	BIGINT	ANO	NE	ANO	
ngram_size	TINYINT	ANO	NE	ANO	
ngram	VARCHAR(150)	NE	NE	ANO	

analyzer_dictionaryUnsortedCZ					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
ngram_id	BIGINT	ANO	NE	ANO	
ngram_size	TINYINT	ANO	NE	ANO	
ngram	VARCHAR(150)	NE	NE	ANO	

analyzer_robotsDisallowTemp					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
disallow_id	INT	ANO	NE	ANO	
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
datetime	DATETIME	NE	NE	ANO	

analyzer_ngramsRules					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
rule_id	INT	ANO	NE	ANO	
min_length	BIGINT	ANO	NE	ANO	
cant_be_num	BOOLEAN	NE	NE	ANO	0

analyzer_ngramsSorted					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
rule_id	SMALLINT	ANO	ANO	ANO	
description	SMALLINT	NE	NE	ANO	0
keywords	SMALLINT	NE	NE	ANO	0
title	SMALLINT	NE	NE	ANO	0
h1	SMALLINT	NE	NE	ANO	0
h2	SMALLINT	NE	NE	ANO	0
h3	SMALLINT	NE	NE	ANO	0
h4	SMALLINT	NE	NE	ANO	0
h5	SMALLINT	NE	NE	ANO	0
h6	SMALLINT	NE	NE	ANO	0
a_href_domain	SMALLINT	NE	NE	ANO	0
a_href_url	SMALLINT	NE	NE	ANO	0
a_href_anchor	SMALLINT	NE	NE	ANO	0
a_href_title	SMALLINT	NE	NE	ANO	0
p	SMALLINT	NE	NE	ANO	0
strong	SMALLINT	NE	NE	ANO	0
em	SMALLINT	NE	NE	ANO	0
b	SMALLINT	NE	NE	ANO	0
i	SMALLINT	NE	NE	ANO	0
li	SMALLINT	NE	NE	ANO	0
dd	SMALLINT	NE	NE	ANO	0
dt	SMALLINT	NE	NE	ANO	0
img_src_domain	SMALLINT	NE	NE	ANO	0
img_src_url	SMALLINT	NE	NE	ANO	0
img_src_alt	SMALLINT	NE	NE	ANO	0
img_src_title	SMALLINT	NE	NE	ANO	0
domain	SMALLINT	NE	NE	ANO	0
url	SMALLINT	NE	NE	ANO	0

analyzer_ngramsSortedCZ					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
rule_id	SMALLINT	ANO	ANO	ANO	
description	SMALLINT	NE	NE	ANO	0
keywords	SMALLINT	NE	NE	ANO	0
title	SMALLINT	NE	NE	ANO	0
h1	SMALLINT	NE	NE	ANO	0
h2	SMALLINT	NE	NE	ANO	0
h3	SMALLINT	NE	NE	ANO	0
h4	SMALLINT	NE	NE	ANO	0
h5	SMALLINT	NE	NE	ANO	0
h6	SMALLINT	NE	NE	ANO	0
a_href_domain	SMALLINT	NE	NE	ANO	0
a_href_url	SMALLINT	NE	NE	ANO	0
a_href_anchor	SMALLINT	NE	NE	ANO	0
a_href_title	SMALLINT	NE	NE	ANO	0
p	SMALLINT	NE	NE	ANO	0
strong	SMALLINT	NE	NE	ANO	0
em	SMALLINT	NE	NE	ANO	0
b	SMALLINT	NE	NE	ANO	0
i	SMALLINT	NE	NE	ANO	0
li	SMALLINT	NE	NE	ANO	0
dd	SMALLINT	NE	NE	ANO	0
dt	SMALLINT	NE	NE	ANO	0
img_src_domain	SMALLINT	NE	NE	ANO	0
img_src_url	SMALLINT	NE	NE	ANO	0
img_src_alt	SMALLINT	NE	NE	ANO	0
img_src_title	SMALLINT	NE	NE	ANO	0
domain	SMALLINT	NE	NE	ANO	0
url	SMALLINT	NE	NE	ANO	0

analyzer_ngramsUnsorted					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
rule_id	SMALLINT	ANO	ANO	ANO	
description	SMALLINT	NE	NE	ANO	0
keywords	SMALLINT	NE	NE	ANO	0
title	SMALLINT	NE	NE	ANO	0
h1	SMALLINT	NE	NE	ANO	0
h2	SMALLINT	NE	NE	ANO	0
h3	SMALLINT	NE	NE	ANO	0
h4	SMALLINT	NE	NE	ANO	0
h5	SMALLINT	NE	NE	ANO	0
h6	SMALLINT	NE	NE	ANO	0
a_href_domain	SMALLINT	NE	NE	ANO	0
a_href_url	SMALLINT	NE	NE	ANO	0
a_href_anchor	SMALLINT	NE	NE	ANO	0
a_href_title	SMALLINT	NE	NE	ANO	0
p	SMALLINT	NE	NE	ANO	0
strong	SMALLINT	NE	NE	ANO	0
em	SMALLINT	NE	NE	ANO	0
b	SMALLINT	NE	NE	ANO	0
i	SMALLINT	NE	NE	ANO	0
li	SMALLINT	NE	NE	ANO	0
dd	SMALLINT	NE	NE	ANO	0
dt	SMALLINT	NE	NE	ANO	0
img_src_domain	SMALLINT	NE	NE	ANO	0
img_src_url	SMALLINT	NE	NE	ANO	0
img_src_alt	SMALLINT	NE	NE	ANO	0
img_src_title	SMALLINT	NE	NE	ANO	0
domain	SMALLINT	NE	NE	ANO	0
url	SMALLINT	NE	NE	ANO	0

analyzer_ngramsUnsortedCZ					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
rule_id	SMALLINT	ANO	ANO	ANO	
description	SMALLINT	NE	NE	ANO	0
keywords	SMALLINT	NE	NE	ANO	0
title	SMALLINT	NE	NE	ANO	0
h1	SMALLINT	NE	NE	ANO	0
h2	SMALLINT	NE	NE	ANO	0
h3	SMALLINT	NE	NE	ANO	0
h4	SMALLINT	NE	NE	ANO	0
h5	SMALLINT	NE	NE	ANO	0
h6	SMALLINT	NE	NE	ANO	0
a_href_domain	SMALLINT	NE	NE	ANO	0
a_href_url	SMALLINT	NE	NE	ANO	0
a_href_anchor	SMALLINT	NE	NE	ANO	0
a_href_title	SMALLINT	NE	NE	ANO	0
p	SMALLINT	NE	NE	ANO	0
strong	SMALLINT	NE	NE	ANO	0
em	SMALLINT	NE	NE	ANO	0
b	SMALLINT	NE	NE	ANO	0
i	SMALLINT	NE	NE	ANO	0
li	SMALLINT	NE	NE	ANO	0
dd	SMALLINT	NE	NE	ANO	0
dt	SMALLINT	NE	NE	ANO	0
img_src_domain	SMALLINT	NE	NE	ANO	0
img_src_url	SMALLINT	NE	NE	ANO	0
img_src_alt	SMALLINT	NE	NE	ANO	0
img_src_title	SMALLINT	NE	NE	ANO	0
domain	SMALLINT	NE	NE	ANO	0
url	SMALLINT	NE	NE	ANO	0

analyzer_parser_a					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
a_id	SMALLINT	ANO	NE	ANO	
href	VARCHAR(100)	NE	NE	NE	
relative	BOOLEAN	NE	NE	ANO	0
anchor	VARCHAR(100)	NE	NE	NE	
image	BOOLEAN	NE	NE	ANO	0
title	VARCHAR(100)	NE	NE	NE	

analyzer_parser_b					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
b_id	SMALLINT	ANO	NE	ANO	
b	TINYTEXT	NE	NE	NE	

analyzer_parser_dd					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
dd_id	SMALLINT	ANO	NE	ANO	
dd	TEXT	NE	NE	NE	

analyzer_parser_description					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
description	TINYTEXT	NE	NE	NE	

analyzer_parser_dt					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
dt_id	SMALLINT	ANO	NE	ANO	
dt	TINYTEXT	NE	NE	NE	

analyzer_parser_em					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
em_id	SMALLINT	ANO	NE	ANO	
em	TINYTEXT	NE	NE	NE	

analyzer_parser_h1					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
h1_id	SMALLINT	ANO	NE	ANO	
h1	TINYTEXT	NE	NE	NE	

analyzer_parser_h2					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
h2_id	SMALLINT	ANO	NE	ANO	
h2	TINYTEXT	NE	NE	NE	

analyzer_parser_h3					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
h3_id	SMALLINT	ANO	NE	ANO	
h3	TINYTEXT	NE	NE	NE	

analyzer_parser_h4					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
h4_id	SMALLINT	ANO	NE	ANO	
h4	TINYTEXT	NE	NE	NE	

analyzer_parser_h5					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
h5_id	SMALLINT	ANO	NE	ANO	
h5	TINYTEXT	NE	NE	NE	

analyzer_parser_h6					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
h6_id	SMALLINT	ANO	NE	ANO	
h6	TINYTEXT	NE	NE	NE	

analyzer_parser_i					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
i_id	SMALLINT	ANO	NE	ANO	
i	TINYTEXT	NE	NE	NE	

analyzer_parser_img					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
img_id	SMALLINT	ANO	NE	ANO	
src	VARCHAR(100)	NE	NE	NE	
relative	BOOLEAN	NE	NE	ANO	
alt	VARCHAR(100)	NE	NE	NE	
title	VARCHAR(100)	NE	NE	NE	

analyzer_parser_keywords					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
keywords	TINYTEXT	NE	NE	NE	

analyzer_parser_li					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
li_id	SMALLINT	ANO	NE	ANO	
li	TINYTEXT	NE	NE	NE	

analyzer_parser_p					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
p_id	SMALLINT	ANO	NE	ANO	
p	TINYTEXT	NE	NE	NE	

analyzer_parser_strong					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
strong_id	SMALLINT	ANO	NE	ANO	
strong	TINYTEXT	NE	NE	NE	

analyzer_parser_keywords					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
title	VARCHAR(100)	NE	NE	NE	

analyzer_recalculatedNgramsSorted					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
weight_id	INT	ANO	ANO	ANO	
count	SMALLINT	NE	NE	NE	

analyzer_recalculatedNgramsSortedCZ					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
weight_id	INT	ANO	ANO	ANO	
count	SMALLINT	NE	NE	NE	

analyzer_recalculatedNgramsUnsorted					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
weight_id	INT	ANO	ANO	ANO	
count	SMALLINT	NE	NE	NE	

analyzer_recalculatedNgramsUnsortedCZ					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
ngram_id	BIGINT	ANO	ANO	ANO	
ngram_size	TINYINT	ANO	ANO	ANO	
weight_id	INT	ANO	ANO	ANO	
count	SMALLINT	NE	NE	NE	

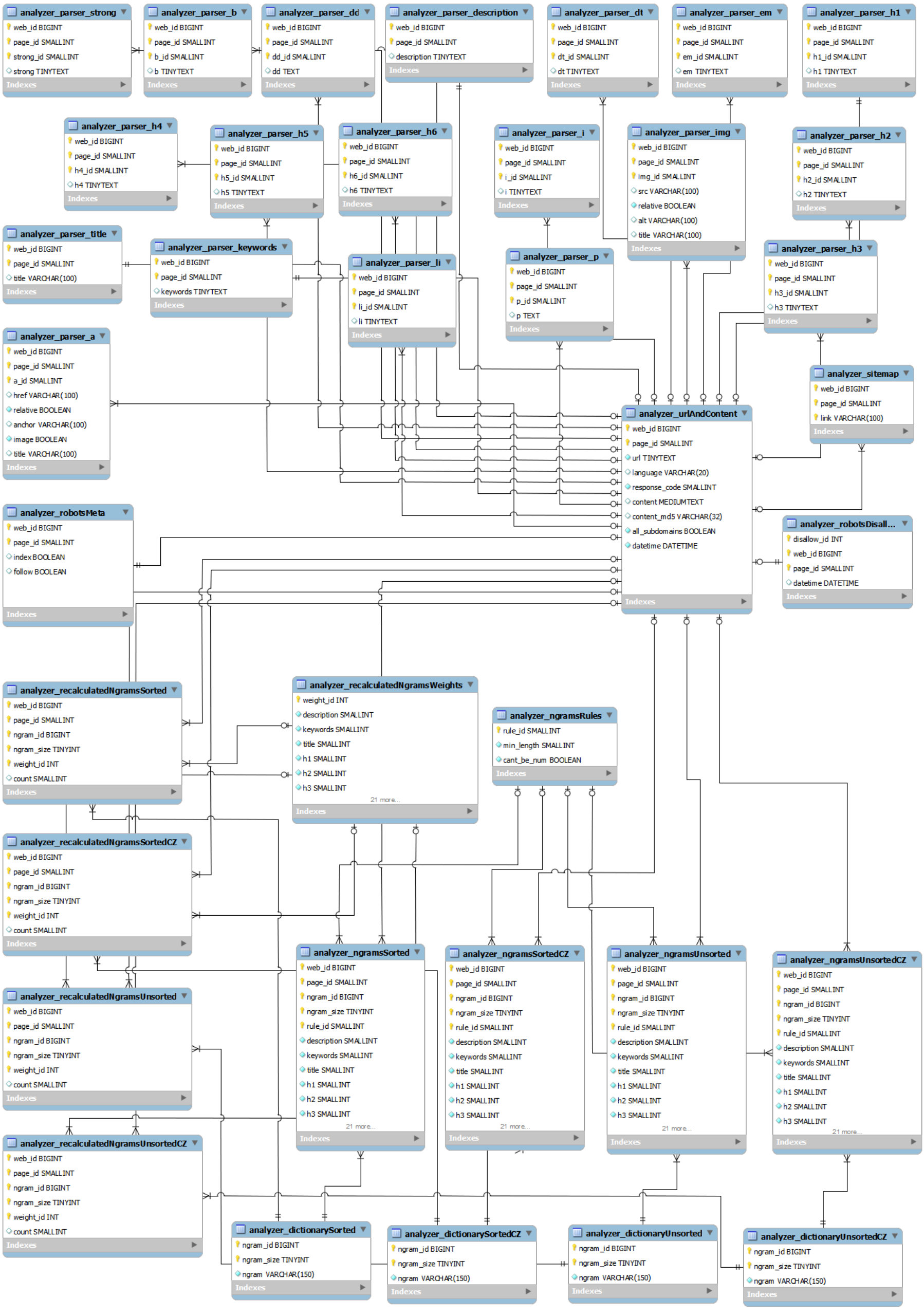
analyzer_ngramsUnsorted					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
weight_id	INT	ANO	NE	ANO	
description	SMALLINT	NE	NE	ANO	0
keywords	SMALLINT	NE	NE	ANO	0
title	SMALLINT	NE	NE	ANO	0
h1	SMALLINT	NE	NE	ANO	0
h2	SMALLINT	NE	NE	ANO	0
h3	SMALLINT	NE	NE	ANO	0
h4	SMALLINT	NE	NE	ANO	0
h5	SMALLINT	NE	NE	ANO	0
h6	SMALLINT	NE	NE	ANO	0
a_href_domain	SMALLINT	NE	NE	ANO	0
a_href_url	SMALLINT	NE	NE	ANO	0
a_href_anchor	SMALLINT	NE	NE	ANO	0
a_href_title	SMALLINT	NE	NE	ANO	0
p	SMALLINT	NE	NE	ANO	0
strong	SMALLINT	NE	NE	ANO	0
em	SMALLINT	NE	NE	ANO	0
b	SMALLINT	NE	NE	ANO	0
i	SMALLINT	NE	NE	ANO	0
li	SMALLINT	NE	NE	ANO	0
dd	SMALLINT	NE	NE	ANO	0
dt	SMALLINT	NE	NE	ANO	0
img_src_domain	SMALLINT	NE	NE	ANO	0
img_src_url	SMALLINT	NE	NE	ANO	0
img_src_alt	SMALLINT	NE	NE	ANO	0
img_src_title	SMALLINT	NE	NE	ANO	0
domain	SMALLINT	NE	NE	ANO	0
url	SMALLINT	NE	NE	ANO	0

analyzer_robotsMeta					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
index	BOOLEAN	NE	NE	NE	
follow	BOOLEAN	NE	NE	NE	

analyzer_sitemap					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	ANO	ANO	
page_id	SMALLINT	ANO	ANO	ANO	
link	VARCHAR(100)	ANO	NE	ANO	

analyzer_urlAndContent					
Atribut	Datový typ	Primární klíč	Cizí klíč	Povinný	Výchozí hodnota
web_id	BIGINT	ANO	NE	ANO	
page_id	SMALLINT	ANO	NE	ANO	
url	TINYTEXT	NE	NE	ANO	
language	VARCHAR(20)	NE	NE	NE	
response_code	SMALLINT	NE	NE	ANO	
content	MEDIUMTEXT	NE	NE	NE	
content_md5	VARCHAR(32)	NE	NE	NE	
all_subdomains	BOOLEAN	NE	NE	ANO	
datetime	DATETIME	NE	NE	ANO	

7.3 E-R Diagram



7.4 Diagram aktivit